Molecular dynamics simulations of haloalkane dehalogenase

Linssen, A.B M

Molecular Dynamics Simulations of Haloalkane Dehalogenase

A statistical analysis of protein motions

RIJKSUNIVERSITEIT GRONINGEN

Molecular Dynamics Simulations of Haloalkane Dehalogenase

A statistical analysis of protein motions

Proefschrift
ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, Dr. D.F.J. Bosscher,
in het openbaar te verdedigen op
vrijdag 20 november 1998
om 13.15 uur

door
Antonius Bernardus Maria Linssen
geboren op 6 mei 1964
te Amstenrade

Promotor: Prof. Dr. H.J.C. Berendsen

Voorwoord

Na zeven jaar is het er dan toch van gekomen. Er is mij in het verleden regelmatig gevraagd wanneer ik zou promoveren en meestal probeerde ik die vraag te ontwijken.

Ik wil dit voorwoord dan ook beginnen met het bedanken van die mensen wiens geduld ik het meest op de proef heb gesteld. Ik denk dan in de eerste plaats aan Jan en Loes. Tijdens het sollicitatiegesprek was ik toch iets te optimistisch over de snelle afronding van m'n promotie. Het moet voor jullie een opluchting zijn geweest om te horen dat de promotiedatum vast stond. Ook Herman Berendsen, mijn promotor, moet af en toe getwijfeld hebben aan een goede afloop. In dit geval speelde mijn gebrek aan tact vast ook een belangrijke rol. Ik wil je hier bedanken voor de grote vrijheid die je me hebt gegeven en uiteraard voor het vertrouwen in mijn werk. De paranimfen, Ulco en Herman, die zich vier jaar geleden al verheugden op het organiseren van een (aardig) feestje, begonnen zich af te vragen wanneer ik eindelijk van hun diensten gebruik zou maken. Jullie hebben in ieder geval de tijd gehad om er over na te denken dus ik heb hoge verwachtingen. En tenslotte natuurlijk mijn familie: Pap, Mam, Jan en Linda, Jos, Gerard, Peter, Hein en Louis. De afgelopen twee jaar heb ik jullie steeds verzekerd dat het boekje aan het einde van de maand af zou zijn. Maar wees eerlijk, uiteindelijk heb ik toch gelijk gekregen.

En dan zijn er mijn voormalige collega's in Groningen. In the first place there is Andrea. It was you who taught me those basics of physics and mathematics that I needed to obtain a deeper understanding of Molecular Dynamics. I also enjoy and appreciate the friendship with Steve, Danilo and Janez and I want to give special thanks to Fabrizia and Andy-Mark for your warm friendship and hospitality whenever I was in Groningen. Van de mensen in de MD groep wil ik verder met name Peter, Bernard, Frans, Natasha en Marc noemen, die me regelmatig geholpen

hebben met allerhande practische zaken. Het dehalogenase-groepje mag natuurlijk niet ontbreken. Frens, Koen, Joost, Geja, Ivo, Rick, Gerrit en Mariel. Onze samenwerking en dicussies hebben een essentiele bijdrage geleverd aan dit boekje. Jiri, if you wouldn't have been in Groningen, I would never have worked on a cis-trans transition. I owe very much to your knowledge and enthousiasm. Veel stimulans kwam ook van Dick Janssen en Bauke Dijkstra. Jullie interesse, maar vooral jullie kritische blik zorgde ervoor dat ik op het juiste spoor bleef.

Veel waardering heb ik ook voor de prettige werksfeer in Utrecht. Behalve Jan en Loes zijn dat bij kristallografie: Stepane, Jan (Kanters), Piet, Erik, de beide Martins, Jean, Anne, Jeroen, Raimond, Roeland, Sjors, Lucy, Carien, Clasien, Arie, Toine, Wijnand, Bouke, Ton, Barend, Marjan, Huub en Roelie. Bij de IR-groep: Joop, Bert, Peter, Hendrik-Jan, Willem, en Toos (ook jullie geduld is overigens ook vaak op de proef gesteld).

Buiten mijn werk zijn er alle vrienden die ik regelmatig verveeld heb met (onder andere) verhalen over m'n onderzoek. Hier wil ik speciaal Leen en Fianne en (nogmaals) Herman en Ulco noemen, bij wie ik ook zo nu en dan mocht overnachten wanneer ik in Groningen was.

Rest mij nog te zeggen dat ik hoop dat we er met z'n allen een leuk feest van weten te maken op 20 november.

Contents

Chapter 1

Introduction

In the course of this century, industrial activities have created great environmental problems by the production of large quantities of chemical waste. Especially during the last decades awareness of the subject has grown and dumping of possibly hazardous compounds is strongly reduced. But the environment is still suffering from a heritage of xenobiotic substances from the past. One way of attacking the problem is by the use of micro-organisms. One such organism is the bacterium *Xanthobacter autotropicus* GJ10 [1]. This bacterium is capable of growing on 1,2-dichloroethane, by using this compound as its sole source of energy and carbon. One of the key enzymes involved in the degradation is *haloalkane dehalogenase* [2] a 35 kd (310 residues) protein which catalyses the reaction:

$$ClCH_2CH_2\boldsymbol{Cl} + H_2O \rightarrow ClCH_2CH_2\boldsymbol{OH} + H^+ + Cl^- \qquad (1.1)$$

Apart from 1,2-dichloroethane, dehalogenase has been shown to catalyse the dehalogenation of a large variety of alkylhalides [2], haloalcohols and halonitriles [4] although it has a low affinity towards these substrates.

This thesis deals with a computational approach to the study of protein behaviour, with haloalkane dehalogenase as the main topic. It is the aim to find a relation between the dynamics of dehalogenase and its

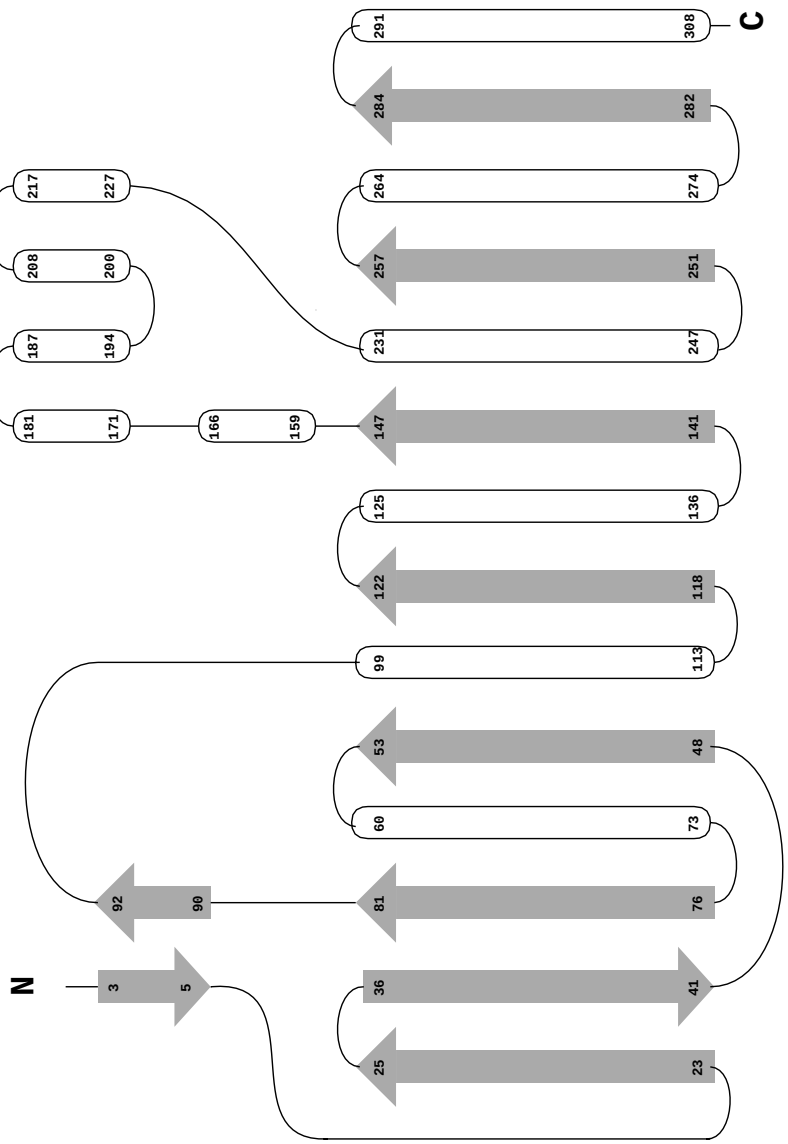

Figure 1.1: Topology of secondary structure elements of dehalogenase [3]

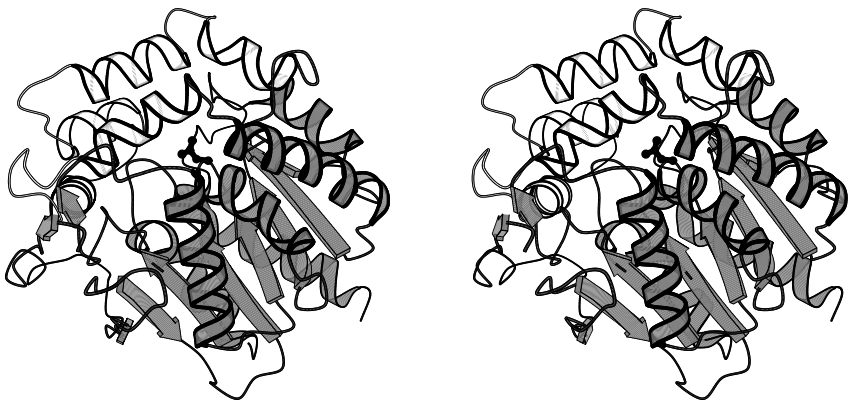

Figure 1.2: Three-dimensional representation of haloalkane dehalogenase. the main domain is drawn in grey, the cap domain in white. In the center, between both domains, the side chain of catalytic residue Asp124 is drawn in black. Drawn with MOLSCRIPT [5]

function. Particular attention will be given to halide release. All the work presented is based on Molecular Dynamics simulations (MD) in combination with some newly developed techniques. Essential Dynamics analysis (ED) [6] will be applied to investigate large globally correlated motions in proteins. Regions that are more or less rigid will be identified with a method referred to as rigid body analysis. To study the effect of a chloride ion bound in the hydrophobic active site cavity, a polarization term has been added to the force field.
All computational methods will be given attention to in chapter 2. Chapter 3 deals in a more elaborate way with the theoretical foundations of ED. In chapters 4 and 5 the results of simulations of dehalogenase respectively with and without a chloride ion in the active site will be presented. Analysis of these results suggest a conformational change of the molecule which will be closely investigated in chapter 6. Finally, in chapter 7, possible consequences of the obtained results for the kinetics of halide release will be discussed.

The X-ray structure of dehalogenase was first solved at a resolution of 2.4 Å [7] and later at 1.9 Å [3]. Fig. 1.2 shows a three-dimensional representation of the backbone structure. The enzyme consists of two domains. The core of the protein is formed by the main domain (residues 1-155 and 230-310) composed of an eight-stranded β-sheet surrounded by six α-helices. On top of it lies the cap domain (residues 156-229), composed of five α-helices (fig. 1.1). It has this topology in common with other hydrolases classified as α/β-hydrolases [8]. A hydrophobic cavity, situated between both domains, forms the active site. By means of X-ray experiments, the structures of the enzyme-substrate complex, a covalently bound enzyme-substrate intermediate and an enzyme-chloride complex were elucidated [9]. Site-directed mutagenesis studies [10, 11, 12, 13] have revealed that Asp124, Trp125, Trp175 and His289 play a role in catalysis. These results have firmly established a two-step catalytic mechanism, shown in fig. 1.3. After formation of the enzyme-substrate complex a nucleophilic attack of Asp124 on the substrate yields a covalently bound intermediate (step 1). The halide ion produced during this reaction is bound between Trp125 and Trp175. As the next step the intermediate is hydrolysed by a water molecule which is activated by His289 (step 2). In the crystal environment, catalysis still occurs, but no complex between the alcohol product and the enzyme was found [9]. The presence of an enzyme-chloride complex suggested that chloride release was the rate-limiting step of the overall reaction. Halide release has been extensively investigated using stopped flow fluorescence experiments [14]. These studies suggested that halide binding can occur via two distinct routes. In one route, fast halide binding is preceeded by a slow enzyme isomerization which was suggested to be a conformational change. In the other route rapid formation of an enzyme-halide collision complex was followed by a slow isomerization. Difference in kinetics between choride and bromide binding were also observed. Fig 1.4 shows the proposed chloride binding scheme. Kinetic constants are listed in table 1.1. E_I and E_{II} represent both enzyme conformations whereas, $E_I.X$ and $E_{II}.X$ their corresponding complexes with the chloride ion. The idea of a con-

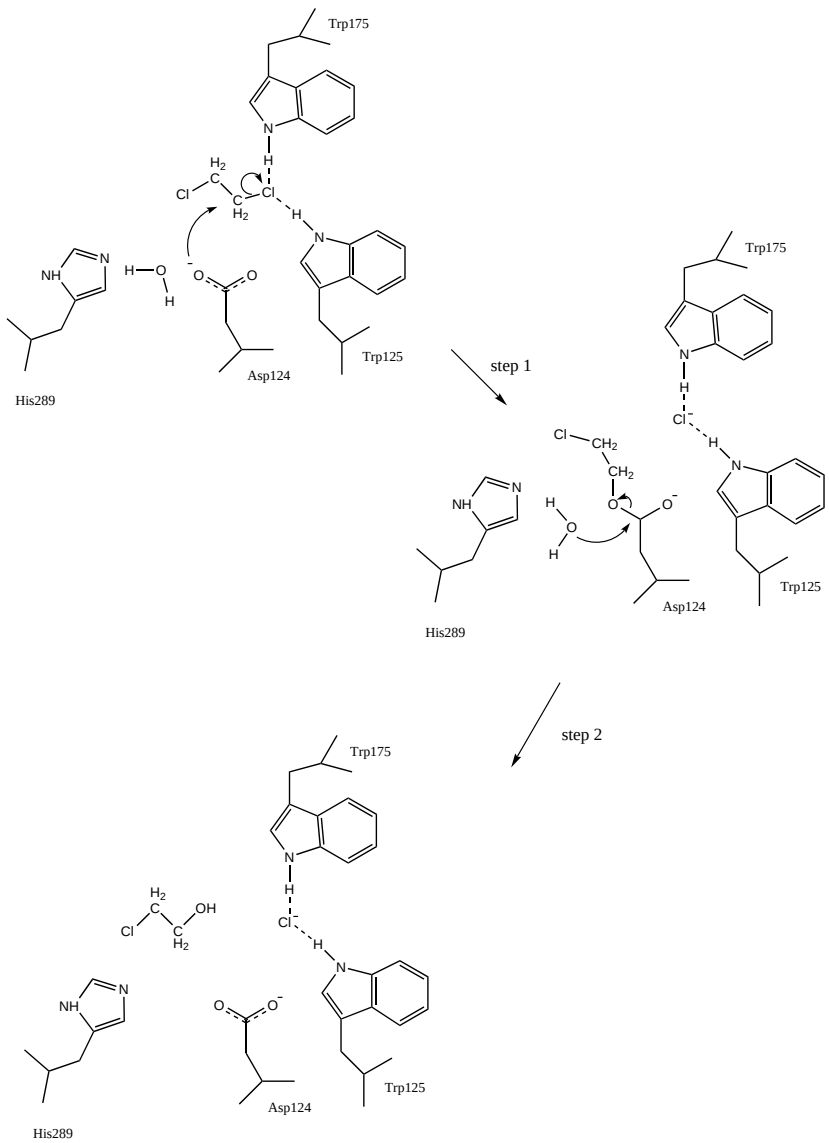

Figure 1.3: Proposed reaction mechanism for dehalogenation of 1,2-dichloroethane (see text).

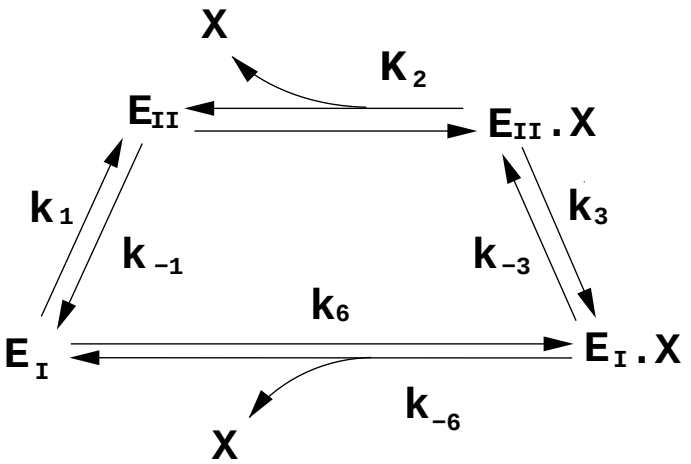

Figure 1.4: Scheme showing both pathways for chloride binding/release [14]

formational change was also supported by a considerable ^2H_2O isotope effect on k_3/ k_{-3} and k_1/ k_{-1} for both chloride as well as bromide release [14].

k_1	$3\pm0.3\ s^{-1}$
k_{-1}	$>300\ s^{-1}$
k_3	$>1450\ s^{-1}$
k_{-3}	$14.5\pm0.5\ s^{-}1$
k_6	$0.0085\pm0.005\ mM^{-1}.s^{-1}$
k_{-6}	$0.66\pm0.03\ s^{-1}$
K_2	-

Table 1.1: kinetic data for fig . 1.4 [14]

Chapter 2

Computational and Theoretical Methods

2.1 The technique of Molecular Dynamics

One way of studying the dynamic behaviour of molecular systems is to simulate their motions by computational techniques. This thesis deals with Molecular Dynamics (MD) simulations that were performed with the GROMOS87 software package [15]. In short one can say that MD is based on Newton's second law:

$$\boldsymbol{F}_i = m_i \boldsymbol{a}_i \tag{2.1}$$

where $\boldsymbol{F}_i$ is the force acting on particle i and $\boldsymbol{a}_i$ is its acceleration. So if at time t_0 the positions $x_i(t_0)$ and the velocities $v_i(t_0)$ are known, one can calculate $x_i(t_0 + \delta t)$ and $v_i(t_0 + \delta t)$, where δt is chosen very small. Knowing the new positions and velocities, the procedure can be repeated over and over again. In this way a trajectory is obtained. The crucial step is in fact the calculation of the forces. The force $\boldsymbol{F}_i$ at time t is determined by the potential energy function, V according to:

$$\boldsymbol{F}_i = -\nabla_i V \tag{2.2}$$

where $\nabla_i V$ is the gradient of the potential energy V with respect to the position coordinates of particle i. In Gromos the potential energy can be calculated as the sum of six terms:

$$V(\boldsymbol{x}) = V(\boldsymbol{x})_{es} + V(\boldsymbol{x})_{lj} + V(\boldsymbol{x})_{bond} + V(\boldsymbol{x})_{angle} + V(\boldsymbol{x})_{imp} + V(\boldsymbol{x})_{dih} \tag{2.3}$$

$V(\boldsymbol{x})_{es}$ is a term describing the Coulombic interaction between electrostatic charges. $V(\boldsymbol{x})_{lj}$ is called the Lennard Jones interaction energy. It reflects the van der Waals interaction between atoms. $V(\boldsymbol{x})_{bond}$ and $V(\boldsymbol{x})_{angle}$ represent the bond stretching and bond angle potential energy. $V(\boldsymbol{x})_{imp}$ is introduced to keep groups of atoms into a spatial (planar or tetrahedral) configuration. Finally $V(\boldsymbol{x})_{dih}$ is a dihedral torsional angle potential. Reference [15] gives a more detailed description of each term. The magnitude of the timestep δt is limited by the highest frequency motions in the system, which are the bond-stretching motions. For this reason bond lengths are often kept rigid by the application of a bond length constraint algorithm (SHAKE, [15, 16]), increasing the allowed timestep δt typically by a factor of four.

2.2 Non-mutual polarization

The force field as described in the first paragraph assumes that charges on the atoms remain unchanged during the simulation. An important term that is missing in this assumption is the polarizability. If an atom or molecule finds itself in the vicinity of an electric charge, its electron cloud will be distorted. As a result, a dipole moment will be induced. In existing force fields, the nonbonded interaction parameters have been adjusted in such a way that polarization effects are, to a certain extent, included. Such adjustments are only valid for the average environment for which the adjustment is designed. But in extreme cases, like for instance an ion in a hydrophobic environment they are not valid. In recent years, algorithms have been developed to incorporate polarizability efficiently into the existing MD algorithms [17, 18, 19]. A review of polarizable water models is given in [20]. But as far as proteins are concerned, we are still far from their implementation. This is mainly due to the fact that not only a set of parameters connected with polarization itself has to be built, but also all other nonbonded interactions have to be reparameterized. This is a task that requires at least several man-years of work.

In our research we were in fact dealing with a chloride ion bound in a hydrophobic environment. With the usual force field this ion is too weakly bound (see chapter 5). Therefore it turned out to be necessary to include, at least in an approximate way, explicit polarization effects. For this reason we introduced a simplified version, referred to as *non-mutual polarization*.

We assumed that the system can be thought of as consisting of two types of particles: *polarizing* and *polarizable* atoms. Polarizing atoms are atoms of which the charge can induce a dipole moment on polarizable atoms. It is also assumed that there is no mutual polarization, i.e. a polarizing atom cannot be polarized itself, and a polarizable atom does not act as a polarizing atom. This appoximation allows a straightforward computation of polarization effects, avoiding the iteration procedure required when mutual polarization is included. When only ions in isolated

positions are considered as polarizing particles, this approximation is quite accurate, but it breaks down for polarized particles in the field of several ions where non-additive effects are not negligeable.

The induced dipole moment $\boldsymbol{\mu}_i$ on polarizable particle P_i at position $\boldsymbol{r}_i$ in the field $\boldsymbol{E}_j$ caused by a polarizing particle P_j at position $\boldsymbol{r}_j$ is given as:

$$\boldsymbol{\mu}_i = \alpha_i \boldsymbol{E}_j, \tag{2.4}$$

where α_i is the polarizability (assumed to be isotropic) of P_i and $\boldsymbol{E}_j$ is given as:

$$\boldsymbol{E}_j = f\frac{q_j}{r^3}\boldsymbol{r}, \tag{2.5}$$

with $f = 1/4\pi\epsilon_0, \boldsymbol{r} = \boldsymbol{r}_i - \boldsymbol{r}_j, r = |\boldsymbol{r}|$ and q_j is the electric charge on P_j. The electric energy of a dipole moment $\boldsymbol{\mu}_i$ in an electric field $\boldsymbol{E}_j$ is given by:

$$V_{el} = -\boldsymbol{\mu}_i \cdot \boldsymbol{E}_j, \tag{2.6}$$

whereas the energy required to polarize P_i is:

$$\begin{aligned} V_{pol} &= \frac{1}{2}\frac{|\boldsymbol{\mu}_i|^2}{\alpha_i} && (2.7)\\ &= \frac{1}{2}\boldsymbol{\mu}_i \cdot \boldsymbol{E}_j && (2.8) \end{aligned}$$

Hence the total contribution of polarization to the potential energy is:

$$\begin{aligned} V_{tot,pol} &= V_{el} + V_{pol} && (2.9)\\ &= -\frac{1}{2}\boldsymbol{\mu}_i \cdot \boldsymbol{E}_j \\ &= -\frac{1}{2}\alpha_i |\boldsymbol{E}_j|^2 \\ &= -\frac{1}{2}\alpha_i f^2 \frac{q_j^2}{r^4} && (2.10) \end{aligned}$$

The contribution to the force exerted on P_i is now given as:

$$\begin{aligned} \boldsymbol{F}_{pol,i}(\boldsymbol{r}) &= -\nabla V_{pol,tot}(\boldsymbol{r}) \\ &= -2\alpha_1 f^2 \frac{q_j^2}{r^6}\boldsymbol{r} && (2.11) \end{aligned}$$

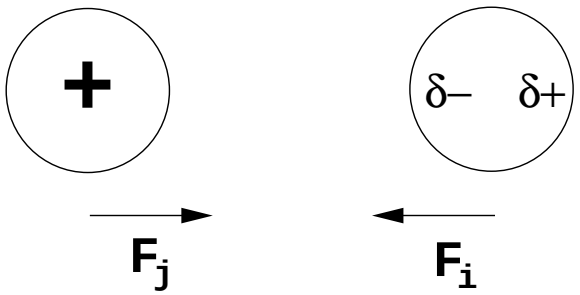

Figure 2.1: If a charged atom j approaches a polarizable atom i a dipole moment is induced on i, resulting in an attractive force between i and j

and the contribution to the force on P_j is:

$$\boldsymbol{F}_{pol,j}(\boldsymbol{r}) = -\boldsymbol{F}_{pol,i}(\boldsymbol{r}) \tag{2.12}$$

If there is no mutual polarizability, the contributions to the force as well as to the potential energy are pair additive with respect to different polarizable atoms. The forces are not pair-additive to polarizing atoms, but since in our application only one polarizing atom is present, this is of no consequence.

In practice, the application of this type of polarization results in additional forces between polarizing and polarizable atoms that are always attractive, as is illustrated in fig 2.1.

Most of the atomic polarizabilities α where taken from Applequist et al. [21]. In their paper they tabulate two types of atomic polarizabilities. One type reproduces (more accurate) molecular polarizabilities with the use of a mutual polarization model. The other type results in (less accurate) molecular polarizabilities by adding all atomic polarizabilities. We chose for the latter, because our model also assumes additivity. Polarizabilities that were not given in [21] were estimated in a way described in table 2.1.

The effect of polarisation can be considerable. The described model was used for simulations of dehalogenase were a chloride ion was bound

in a hydrophobic cavity (chapters 5 and 6). To give a rough estimate of the relative contributions to the energy, we took a structure from the simulation in chapter 5 (the conformation at 20 ps) and calculated the total interaction energies of the chloride ion with its environment. The Lennard-Jones, electrostatic and polarisation energies were respectively 12, -305 and -112 kJ mole^{-1} . This can be compared to a value of +360 kJ mole^{-1} free energy change for the removal of a chloride ion from dehalogase. This value follows from the experimental chloride binding constant (78 mM), the free energy of Cl^- hydration (-317 kJ mole^{-1}), and the standard translational entropy of Cl^- at 1 M concentration. We see that, by using the model described above, the polarisation term adds more than 30% to the stabilization energy, and is essential to keep the chloride in its place.

atom type	α^* (Å^3)
C (aliphatic)	1.027^a
CH (aliphatic)	1.434^a
CH_2 (aliphatic)	1.841^a
CH_3 (aliphatic)	2.248^a
C (carbonyl)	1.027^a
C (aromatic)	1.39^b
H (aliphatic)	0.407^a
H (aromatic)	0.32^b
H (amide,peptide,pyrrole)	0.23^c
O (carbonyl)	0.841^a
N (amide,peptide,pyrrole)	1.40^c
a) taken from ref. [21]	
b) derived from the polarizabilities of benzene(C_6H_6) and fluorobenzene (C_6H_5F) from ref. [22] where the atomic polarizability of F was taken from ref.[21].	
c) derived from the polarizabilities of trimethylamine ($N(CH_3)_3$) and NH_3 from ref. [22] where the polarizability of methyl groups was taken from [21]	

Table 2.1: Atomic polarizabilities α^* used for the non-mutual polarizability model. Note that $\alpha^* = \alpha f$ (which in eqs. 2.10 and 2.11, using SI units, should be expressed in m^3) is used in the tabulated values.

2.3 Essential Dynamics: the basic concepts

2.3.1 Summary

In this section a simplified description is given of the essential dynamics method that is fully described in chapter 3. We consider two special cases to demonstrate the principles of the method.

2.3.2 Correlations between fluctuations

One of the major problems in interpreting MD results is the large number of coordinates that determine the configuration of a system. When one considers a small protein of about 1000 atoms ($\sim$100 aminoacid residues), there will be 3000 position coordinates to take care of. A number that doesn't even include solvent molecules, which have a large effect on a protein's behaviour. Fortunately, not all combinations of coordinates are allowed. For instance, atoms that are interconnected by a chemical bond will always be at an almost constant distance with respect to each other. Similarly, bond angles have a restricted freedom. Also the secondary and even the tertiary structure will reduce the number of possible coordinate combinations. If there would be no mutual relationships between the atoms, we would not deal with a protein but with an ideal gas.

These correlations between coordinates can be used to reduce the number of *degrees of freedom* that are necessary to describe the configuration of a system. The most straightforward example of this is a two-dimensional particle P, that can only move along a straight the line $x = y$ (fig. 2.2). There exists a correlation between x and y coordinates such that if one coordinate is known, the other can be derived from it. The problem is how to discover the mathematical description of existing correlations: In the above example it is necessary first to find out that the particle moves in a straight line. This can be done by simulating its motions. It is then clear, by visual inspection, that it moves linearly. For systems of three dimensions, containing more than one particle, with

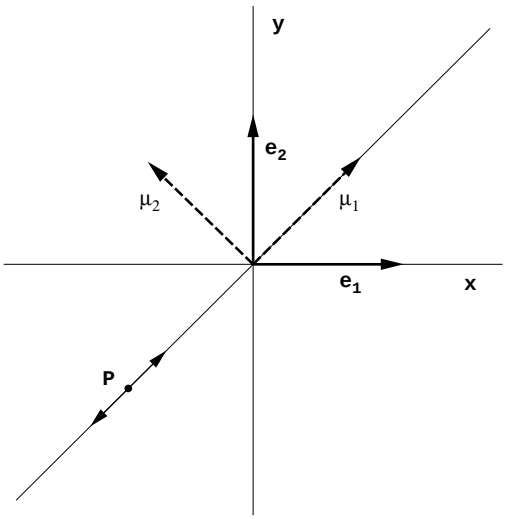

Figure 2.2: Particle P can only move along the line $x = y$

more complicated types of correlation, this method is obviously not very useful.

2.3.3 Reducing the number of coordinates

We will now describe a mathematical method for determining correlations between position coordinates. The best way to do this is to stick with our example of the two-dimensional particle and introduce some basic concepts of linear algebra.
It is general practise to use cartesian coordinate axes, designated as x- and y-axis. If we define two *basis vectors* $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ along these axes, each of unit length, then a point P with coordinates $x = a$ and $y = b$ is

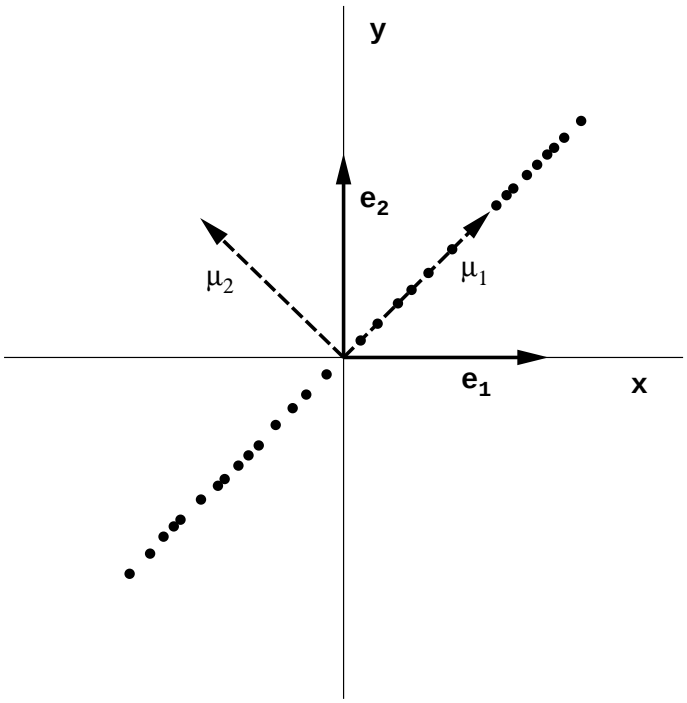

Figure 2.3: Positions of P collected from a simulation

represented by a vector:

$$\boldsymbol{p} = a\boldsymbol{e}_1 + b\boldsymbol{e}_2 \tag{2.13}$$

The numbers a and b, usually arranged as a column matrix $\begin{pmatrix} a \\ b \end{pmatrix}$ are a *representation* of the vector $\boldsymbol{p}$ on the basis set $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$. The representation of $\boldsymbol{e}_1$ is $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and of $\boldsymbol{e}_2$ is $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$

We could also use another set of basis vectors, for instance: $\{\boldsymbol{\mu_1}, \boldsymbol{\mu_2}\}$ given as:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \boldsymbol{\mu}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \tag{2.14}$$

or rather:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} \frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} \end{pmatrix} \quad \boldsymbol{\mu}_2 = \begin{pmatrix} -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} \end{pmatrix} \tag{2.15}$$

because, mathematically, it is more convenient to deal with normalized basis vectors (i.e. basis vectors of unit length). The position of P is now given as:

$$\alpha\boldsymbol{\mu}_1 + \beta\boldsymbol{\mu}_2 \tag{2.16}$$

It can easily be verified that $\alpha = a\sqrt{2}$ and, more interesting, $\beta = 0$ for all positions of P. This means that if the set of vectors $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$ is chosen as a basis, only $\boldsymbol{\mu}_1$, with its coordinate α, is required to determine the exact position of P.

We now turn to the question how to obtain the set $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$. For this we assume that the motion of P has been simulated over a certain time interval, and that during this simulation positions have been sampled and drawn in a graph (fig. 2.3). If we have N positions P_i (and vectors $\boldsymbol{p}_i$), where $i = 1, 2, ...N$, each with cartesian coordinates (a_i, b_i), these positions can be used to construct a *correlation* or *covariance* matrix C defined as:

$$C = \begin{pmatrix} \langle (a - \langle a \rangle)^2 \rangle & \langle (a - \langle a \rangle)(b - \langle b \rangle) \rangle \\ \langle (b - \langle b \rangle)(a - \langle a \rangle) \rangle & \langle (b - \langle b \rangle)^2 \rangle \end{pmatrix} \tag{2.17}$$

where $\langle\rangle$ indicate averages. For the sake of clarity we have chosen the positions in fig.2.3 such that $\langle a \rangle = \langle b \rangle = 0$. In this way the expression for C reduces to:

$$C = \begin{pmatrix} \langle a^2 \rangle & \langle ab \rangle \\ \langle ba \rangle & \langle b^2 \rangle \end{pmatrix} \tag{2.18}$$

Now we have to keep in mind that we are dealing with cartesian coordinates this means that a_i and b_i are the coordinates of P_i with respect to $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$. We could also construct a covariance matrix based on coordinates (α_i, β_i), expressed in $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. This would result in (remembering that $\beta = 0$ for all positions):

$$\Lambda = \begin{pmatrix} \langle 2\alpha^2 \rangle & 0 \\ 0 & 0 \end{pmatrix} \tag{2.19}$$

The entries of Λ can be split into diagonal (from top left to bottom right) and off-diagonal elements. The latter are all zero, which means that, in mathematical terms, we have *diagonalized* C by changing the basis from $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$ to $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$. The new basis vectors are called eigenvectors of C. The diagonal elements are called eigenvalues, and each one has an eigenvector associated with it. In our case we have eigenvalues $\lambda_1 = \langle 2\alpha^2 \rangle$ and $\lambda_2 = 0$ corresponding to respectively $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. The eigenvalues are in fact the *mean square displacements* (msd) of P in the direction of the corresponding eigenvector. If the eigenvectors are normalized, the transformation from the cartesian coordinates (a_i, b_i) of a point P_i to the new ones (α_i, β_i) can easily be accomplished by calculating *inner products*:

$$\begin{aligned} \alpha_i &= \boldsymbol{\mu}_1 \cdot \boldsymbol{p}_i \\ &= \tfrac{1}{2}\sqrt{2}a_i + \tfrac{1}{2}\sqrt{2}b_i \\ &= \sqrt{2}a_i \end{aligned} \qquad \begin{aligned} \beta_i &= \boldsymbol{\mu}_2 \cdot \boldsymbol{p}_i \\ &= -\tfrac{1}{2}\sqrt{2}a_i + \tfrac{1}{2}\sqrt{2}b_i \\ &= 0 \end{aligned} \tag{2.20}$$

where we have used the fact that $a_i = b_i$. α_i and β_i are called *projections* of $\boldsymbol{p}$ on $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ respectively.

2.3.4 A multidimensional example

The method described above can also be applied to systems of more then one particle. To demonstrate this, we will use a model [1] of a mechanical scissor-like construct as shown in fig. 2.4. We have seven particles $P_1, P_2, ...P_7$, that are connected to their neighbouring particle with rigid bonds of one unit length. The particles themselves act as hinges so as to permit the system to change its *configuration* as shown. The configuration of the system is only dependent on the angle ϕ between bonds $P_1 - P_2$ and $P_1 - P4$. Now we will change the angle, at constant velocity, within one time unit from $\phi(t) = 0$ at $t = 0$ to $\phi = 180^o$ at $t = 1$. Thus,

[1] This example was suggested by Dr. H. Bekker, Dept. of Computer Science, University of Groningen

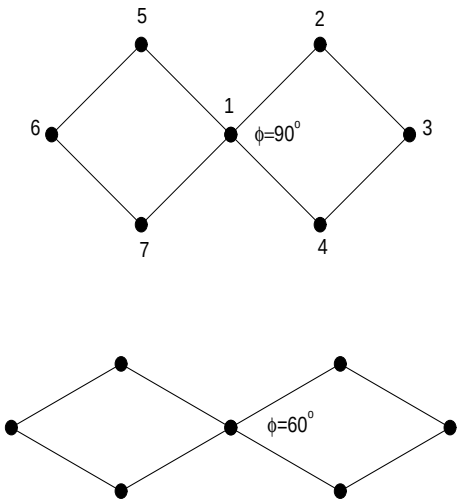

Figure 2.4: Two conformations of the mechanical construct illustrating its mechanism.

at the start, all particles are positioned on the x-axis (fig. 2.5a), when moving they will pass through configurations b and c to finally end at the positions of structure d. As we have seven particles, there are fourteen coordinates at each time t, they will be called: $x_1(t), x_2(t), ...x_{14}(t)$ (x_1 refers to the x-coordinate of P_1, x_2 to the y-coordinate of P_1, x_3 to the x-coordinate of P_2,...etc.). The configuration at time t can be represented as a fourteen-dimensional vector $\boldsymbol{x}(t)$:

$$\boldsymbol{x}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_{14}(t) \end{pmatrix} \tag{2.21}$$

Unlike the previous example the average value of most of the coordinates will not be zero, so they will also have to be included into the covariance matrix. C will be a 14×14 -dimensional matrix. In analogy with eq. 2.17, the entry in the i^{th} row and j^{th} column of C is given as:

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \tag{2.22}$$

Equation. 2.17 shows C.

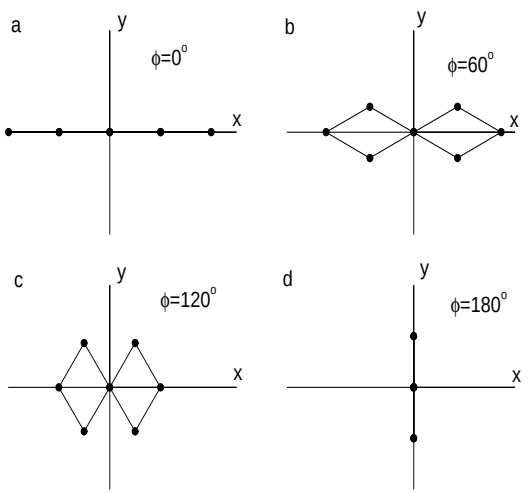

Figure 2.5: Four conformations that are adapted when going from $\phi = 0$ to $\phi = 180$.

$$
C = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.09 & -0.09 & 0.19 & 0 & 0.09 & -0.09 & 0.09 & -0.09 & -0.19 & 0 & -0.09 & 0.09 & 0.09 \\
0 & 0 & -0.09 & 0.09 & -0.17 & 0 & -0.09 & 0.09 & -0.09 & 0.09 & 0.17 & 0 & 0.09 & -0.09 & -0.09 \\
0 & 0 & 0.19 & -0.17 & 0.38 & 0 & 0.19 & 0.17 & -0.19 & 0.17 & -0.38 & 0 & -0.19 & 0.17 & 0.17 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.09 & -0.09 & 0.19 & 0 & 0.09 & 0.09 & -0.09 & -0.9 & -0.19 & 0 & -0.09 & 0.09 & 0.09 \\
0 & 0 & 0.09 & -0.09 & 0.17 & 0 & 0.09 & 0.09 & -0.09 & -0.09 & -0.17 & 0 & -0.09 & 0.09 & 0.09 \\
0 & 0 & -0.09 & 0.09 & -0.19 & 0 & -0.09 & -0.09 & 0.09 & 0.09 & 0.019 & 0 & 0.09 & -0.09 & -0.09 \\
0 & 0 & -0.09 & 0.09 & -0.17 & 0 & -0.09 & -0.09 & 0.09 & 0.09 & 0.17 & 0 & 0.09 & -0.09 & -0.09 \\
0 & 0 & -0.19 & 0.17 & -0.38 & 0 & -0.19 & -0.17 & 0.19 & 0.17 & 0.38 & 0 & 0.19 & -0.17 & -0.17 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -0.09 & 0.09 & -0.19 & 0 & -0.09 & -0.09 & 0.09 & 0.09 & 0.19 & 0 & 0.09 & -0.09 & -0.09 \\
0 & 0 & 0.09 & -0.09 & 0.17 & 0 & 0.09 & 0.09 & -0.09 & -0.09 & -0.17 & 0 & -0.09 & 0.09 & 0.09
\end{pmatrix} \tag{2.23}
$$

Diagonalizing C results in two nonzero eigenvalues, namely: $\lambda_1 = 1.469$ and $\lambda_2 = 0.046$ with corresponding eigenvectors:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0.0 \\ 0.0 \\ -0.253 \\ 0.242 \\ -0.505 \\ 0.0 \\ -0.253 \\ -0.242 \\ 0.253 \\ 0.242 \\ 0.505 \\ 0.0 \\ 0.253 \\ -0.242 \end{pmatrix} \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 0.0 \\ 0.0 \\ 0.140 \\ 0.438 \\ 0.279 \\ 0.0 \\ 0.140 \\ -0.438 \\ -0.140 \\ 0.438 \\ -0.279 \\ 0.0 \\ -0.140 \\ -0.438 \end{pmatrix} \tag{2.24}$$

In total C has fourteen eigenvectors, but because twelve of them have eigenvalues that are equal to zero, in the fourteen-dimensional space there is no motion along their direction. This means that only the coordinates (or projections) p_1 and p_2, corresponding to $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are required to completely determine the configuration of the system. The projections, as a function of time, can be calculated like eq. 2.20, the only difference being that, also here, the average coordinates enter the equation:

$$\begin{aligned} p_1(t) &= \boldsymbol{\mu_1} \cdot [\boldsymbol{x}(t) - \langle \boldsymbol{x} \rangle] & p_2(t) &= \boldsymbol{\mu_2} \cdot [\boldsymbol{x}(t) - \langle \boldsymbol{x} \rangle] \\ &= \sum_{i=1}^{14} \mu_{1,i}[x_i(t) - \langle x_i \rangle] & &= \sum_{i=1}^{14} \mu_{2,i}[x_i(t) - \langle x_i \rangle] \end{aligned} \tag{2.25}$$

were $\mu_{1,i}$ is the i^{th} coefficient of $\boldsymbol{\mu_1}$ (and likewise for $\mu_{2,i}$) In fig. 2.6 the projections are plotted versus time. The largest displacements are those along $\boldsymbol{\mu_1}$. This is in accordance with the eigenvalue. As was already mentioned, the eigenvalue is the mean square displacement (msd) along the corresponding eigenvector, where the msd is given as the average of the square of the projection:

$$msd = \langle p^2 \rangle \tag{2.26}$$

For reconstruction of the structure at a certain time t, we only need $p_1(t)$ and $p_2(t)$. This is illustrated in the example below were the structure

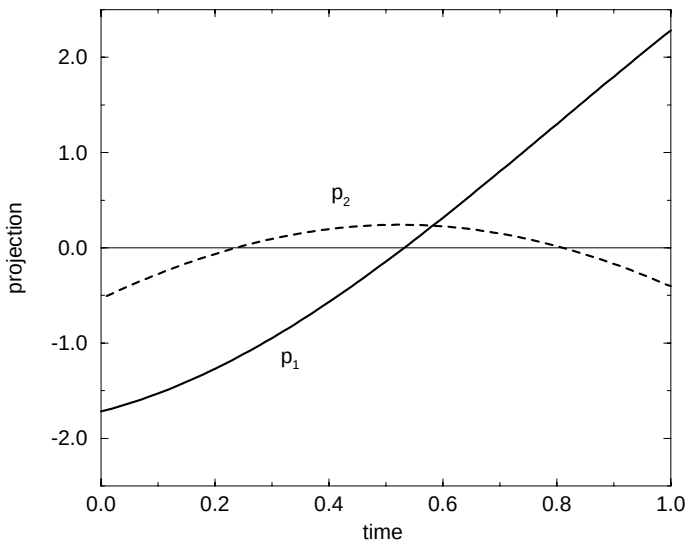

Figure 2.6: Projections p_1 and p_2 along $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ respectively

at $t = 0$ is reconstructed. At time t=0 we have $p_1(0) = -1.717$ and $p_2(0) = -0.505$. The complete structure $\boldsymbol{x(t)}$ is now given as:

$$\boldsymbol{x}(0) = \langle \boldsymbol{x} \rangle + p_1(0)\boldsymbol{\mu_1} + p_2(0)\boldsymbol{\mu_2} \tag{2.27}$$

resulting in:

$$\boldsymbol{x}(0) = \begin{pmatrix} 0.0 \\ 0.0 \\ 0.637 \\ 0.637 \\ 1.273 \\ 0.0 \\ 0.637 \\ -0.637 \\ -0.637 \\ 0.637 \\ -1.273 \\ 0.0 \\ -0.637 \\ -0.637 \end{pmatrix} - 1.717 \begin{pmatrix} 0.0 \\ 0.0 \\ -0.253 \\ 0.242 \\ -0.505 \\ 0.0 \\ -0.253 \\ -0.242 \\ 0.253 \\ 0.242 \\ 0.505 \\ 0.0 \\ 0.253 \\ -0.242 \end{pmatrix} - 0.505 \begin{pmatrix} 0.0 \\ 0.0 \\ 0.140 \\ 0.438 \\ 0.279 \\ 0.0 \\ 0.140 \\ -0.438 \\ -0.140 \\ 0.438 \\ -0.279 \\ 0.0 \\ -0.140 \\ -0.438 \end{pmatrix} = \begin{pmatrix} 0.0 \\ 0.0 \\ 1.0 \\ 0.0 \\ 2.0 \\ 0.0 \\ 1.0 \\ 0.0 \\ -1.0 \\ 0.0 \\ -2.0 \\ 0.0 \\ -1.0 \\ 0.0 \end{pmatrix} \tag{2.28}$$

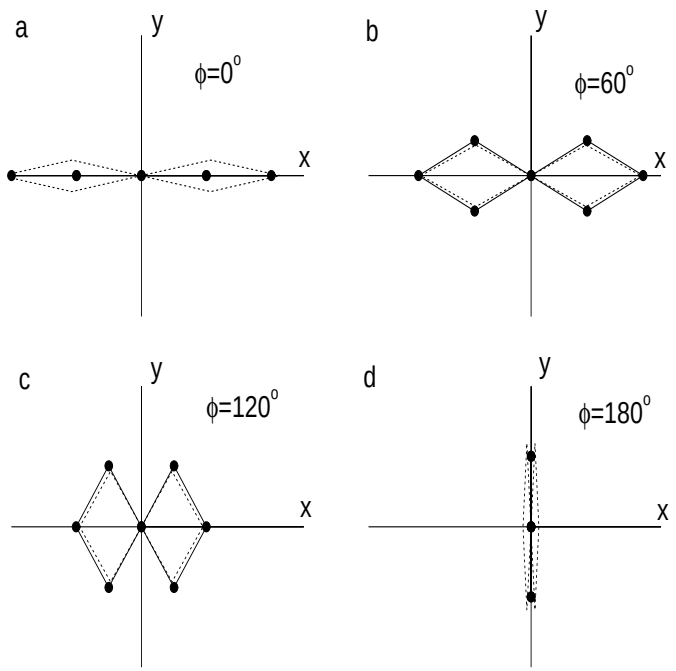

Figure 2.7: The solid line shows the same structures as in fig. 2.6, the dotted line shows these conformations approximated by taking only the projections along $\boldsymbol{\mu_1}$

These are the exact coordinates of structure a in fig. 2.5. Another interesting feature in this example is the fact that λ_1 is significantly larger than λ_2 meaning that most of the motion takes place in the direction of $\boldsymbol{\mu_1}$. This implies that we can approximate the exact structure at time t by only using $p_1(t)$, or:

$$\boldsymbol{x}(t) \approx \langle \boldsymbol{x} \rangle + p_1(t)\boldsymbol{\mu_1} \tag{2.29}$$

In fig. 2.7 we have approximated the structures from fig. 2.5 in this way. It can be seen from this figure that structures at $t = 0$ and $t = 1$ have the largest deviations from the real structure. This fact can be deduced from fig. 2.6, if we note that the displacement of p_2 is largest at $t = 0$ and

$t = 1$. However the overall behaviour of the construct is still very well described. It should be noted that the accuracy of this approximation depends on the magnitude of the eigenvalue corresponding to the eigenvector that is used. If we had used the second eigenvector, the accuracy would have been significantly smaller.

The reader may have noticed that the motion can actually be described by the use of only one variable, namely ϕ. It may seems strange that our analysis results in two variables. To understand this we have to realize that the eigenvectors in equation 2.24 are expressed in cartesian coordinates. It can easily be shown that when motion takes place along one single eigenvector, it has to be linear in three dimensions. It can also be seen that particles 2,4,5 and 7 move in a circular way, so their motion cannot be described by one single vector. This does not imply that the motions along $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are independent, but the dependence is non-linear. If we want to express the full configuration of the system at time t with only one coordinate we obtain the expression:

$$\boldsymbol{x}(t) = \begin{pmatrix} 0.0 \\ 0.0 \\ \cos\phi(t) \\ \sin\phi(t) \\ 2\cos\phi(t) \\ 0.0 \\ \cos\phi(t) \\ -\sin\phi(t) \\ -\cos\phi(t) \\ \sin\phi(t) \\ -2\cos\phi(t) \\ 0.0 \\ -\cos\phi(t) \\ -\sin\phi(t) \end{pmatrix} \tag{2.30}$$

The application of the method to proteins will be the subject of the next chapter, where we will also deal with the more fundamental theoretical concepts.

2.4 Rigid body analysis

After analyzing the essential degrees of freedom of a protein one obtains information about its internal correlated motions. Sometimes these motions are rather complex and it is not always easy to relate them to the function of the protein. Interpretation of the essential motion could be facilitated if one has a clear picture of the architecture of the protein, i.e. if one can identify regions that can be considered as being rigid, and that move with respect to each other. Hayward et al.[23] introduced a method, based on Normal Modes Analysis or Essential Dynamics, in which they assumed that the motions along respectively low frequency modes or large eigenvalue vectors approximate inter-domain motions. Later it was adapted to identify domains in any protein for which more than one conformation is known [24]. Here we present a method that defines rigid bodies based on interatomic distance fluctuation obtained from Molecular Dynamics simulations. It will be referred to as *rigid body analysis*.

The mathematical concepts behind the method were, among others, developed by Hill [25]. Later Holm and Sander [26] exploited them in a powerful method which enabled them to identify structural units in proteins. Holm and Sander based their approach on the interactions of all residue pairs in the molecule. Their aim was to form clusters of atoms in such a way that the interactions within each cluster were maximized, whereas mutual interactions among clusters were minimized. The procedure is based on a statistical analysis of the contact matrix $\boldsymbol{A}$, of which each element A_{ij} is an approximate (positive) value for the strength of the interaction between residue i and j. To obtain the desired clusters, scores x_i are assigned to each residue i, in such a way that, if A_{ij} is the weight of a point (x_i, x_j) in a two-dimensional space, the set of points $\{(x_i, x_j)\}$ have a maximum correlation coefficient. This corresponds to the statistical method of correspondence analysis [25]. Residues can now be clustered: residues with similar scores belong to the same cluster. These scores are determined by solving the eigenvalue

problem:

$$\rho \boldsymbol{x} = (\boldsymbol{R}^{-1}\boldsymbol{A})\boldsymbol{x} \tag{2.31}$$

where $\boldsymbol{R}$ is a diagonal matrix defined as

$$R_{ii} = \sum_j A_{ij} \tag{2.32}$$

Equation 2.31 has one trivial and irrelevant solution, namely $\boldsymbol{x}$=(1,1,...) with eigenvalue 1, which is the maximal eigenvalue. The actual solution that gives the desired scores is the solution corresponding to the second largest eigenvalue. This eigenvalue ρ is the correlation coefficient of the points (x_i, x_j):

$$\rho = \frac{\sum_i \sum_j A_{ij} x_i x_j}{\sum_i \sum_j A_{ij} x_i^2} \tag{2.33}$$

In order to solve equation 2.31, it is first rearranged. From equation 2.31 we obtain:

$$\rho^2 \boldsymbol{x} = (\boldsymbol{R}^{-1}\boldsymbol{A})(\boldsymbol{R}^{-1}\boldsymbol{A})\boldsymbol{x} \tag{2.34}$$

If $\boldsymbol{R}^{\frac{1}{2}}$ is a diagonal matrix defined as $R_{ii}^{\frac{1}{2}} = \sqrt{R_{ii}}$ we may write

$$\rho^2 \boldsymbol{x} = \boldsymbol{R}^{-\frac{1}{2}}\boldsymbol{R}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{R}^{-\frac{1}{2}}\boldsymbol{R}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{R}^{-\frac{1}{2}}\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{x} \tag{2.35}$$

Multiplying both sides with $\boldsymbol{R}^{\frac{1}{2}}$ results in:

$$\rho^2(\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{x}) = (\boldsymbol{R}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{R}^{-\frac{1}{2}})(\boldsymbol{R}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{R}^{-\frac{1}{2}})(\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{x}) \tag{2.36}$$

Rearranging is done because $(\boldsymbol{R}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{R}^{-\frac{1}{2}})(\boldsymbol{R}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{R}^{-\frac{1}{2}})$ is a semi-definite symmetric matrix, which makes equation 2.36 numerically easier to solve.

Given the scores, the molecule is divided into two units. This process is repeated recursively on each newly obtained subdomain until certain limits (for instance a minimum number of residues) are reached. In this way a tree decomposition is obtained.

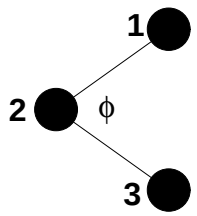

Figure 2.8: triatomic molecule whose only internal degree of freedom is a bending of angle ϕ

Rigid body analysis uses the same mathematical principles. Instead of using the inter-atomic interactions as a criterion to subdivide the molecule we use the inter-atomic distance fluctuations:

$$\langle (d_{ij} - \langle d_{ij} \rangle)^2 \rangle = \langle d_{ij}^2 \rangle - \langle d_{ij} \rangle^2 \tag{2.37}$$

where d_{ij} is the distance between atoms i and j and $\langle\rangle$ indicates averages obtained from the simulation. We could now build a matrix from these fluctuations. We then define the matrix $\boldsymbol{D}$ with elements:

$$D_{ij} = \langle (d - \langle d \rangle)^2 \rangle_{max} - \langle (d_{ij} - \langle d_{ij} \rangle)^2 \tag{2.38}$$

where $\langle (d_{ij} - \langle d_{ij} \rangle)^2 \rangle_{max}$ is the maximum fluctuation among all distances. $\boldsymbol{D}$ contains only positive values, and the largest ones correspond to atom pairs with the smallest inter-atomic distance fluctuations. Rigid regions can be characterized as clusters within which the distance fluctuations are minimal (or D_{ij} is maximal), and between which large fluctuations exist. By substituting $\boldsymbol{A}$ in equation 2.31 with $\boldsymbol{D}$ we can now calculate scores that cluster rigid regions. Interpretation of these scores is however not always trivial as can be illustrated by a simple triatomic molecule for which the angular bending is the only internal degree of freedom (fig. 2.8). This molecule has two rigid regions formed by atoms 1 and 2 and atoms 2 and 3. Atom 2 belongs to both regions. The (normalized) scores that will be found for atoms 1-3 will be $\frac{1}{2}\sqrt{2}, 0$, and $-\frac{1}{2}\sqrt{2}$, respectively. This shows that atoms belonging to the same rigid region do not necessarily

have similar scores. This effect may point towards hinge regions, but the scores do not contain this information. On the other hand it is also possible that during a simulation, two independent rigid regions by chance do not move with respect to each other, and therefore may get the same scores. It is clear that, without any further analysis, visual inspection of the motion remains necessary.

Chapter 3

Essential Dynamics of Proteins[1]

3.1 Introduction

Functional proteins are generally stable mechanical constructs that allow certain types of internal motion to enable their biological function. The internal motions may allow the binding of a substrate or coenzyme, the adaptation to a different environment as in specific aggregation, or the transmission of a conformational adjustment to affect the binding or reactivity at a remote site, as in allosteric effects. Such functional internal motions may be subtle and involve complex correlations between atomic motions, but their nature is inherent in the structure and interactions within the molecule. It is a challenge to derive such motions from the molecular structure and interactions, to identify their functional role, and to reduce the complex protein dynamics to its essential degrees of freedom.

We investigate the correlations between atomic positional fluctuations in a protein, as derived from (nanosecond) molecular dynamics

[1]Most of this chapter was published in: A. Amadei, A.B.M. Linssen and H.J.C. Berendsen, Essential Dynamics of Proteins, *Proteins:* (1993), **17**, 412-425

(MD), both in vacuum and in aqueous environment. By diagonalizing the covariance matrix of the atomic displacements, we find that most of the positional fluctuations are concentrated in correlated motions in a subspace of only a few (not more than 1%) degrees of freedom, while all other degrees of freedom represent much less important, basically independent, Gaussian fluctuations orthogonal to the "essential" subspace. The motion outside the essential subspace can be considered as essentially constrained. This offers the possibility of representing protein dynamics in the essential subspace only.

Our treatment differs from a harmonic or quasi-harmonic normal mode analysis [27, 28, 29, 30, 31] in two ways. First, we do not analyze the motion but rather the positional fluctuations, without involving the atomic masses in the analysis. Our purpose is to identify an "irrelevant" subspace which may be considered essentially constrained. Second, we do not attempt to describe the motion in the "essential" subspace as harmonic, or even as mutually uncoupled, because it is neither harmonic nor uncoupled and such a treatment would restrict the mechanics of a protein to the level of uninteresting vibrations. The projection of a MD trajectory onto normal mode axes as carried out by Horiuchi and Gō [28] bears a resemblance to our analysis of displacements in the essential subspace, be it that the spaces onto which the motion is projected are not the same: they consider a dihedral angle subspace defined by normal modes of low frequency: we retain Cartesian coordinates and define the subspace from the covariance matrix. They find that the motions in the lower modes are restricted in narrower ranges than those derived by the harmonic approximation; we find a similar restriction due to nonlinear behaviour and the presence of nonlinear constraints within the essential subspace. Our analysis is in fact identical to the one described by Garcia [32]. He found that the largest linearly correlated motions in a protein are defined by the eigenvectors of the covariance matrix with the largest eigenvalue. Also he recognized that these motions are far from harmonic. In fact, his and our approach correspond to Principal Component Analysis (PCA) of the configurational space.

The covariance matrix of atomic displacements has been used previously for quasiharmonic analysis [29, 30], for entropy determination [33, 34, 35], and for an analysis of collective motions [27]. We note that Ichiye and Karplus [27] construct the $N \times N$ covariance matrix, where N is the number of atoms in the system, of the vectorial inner products, while we construct the $3N \times 3N$ matrix of Cartesian displacements. The former indicates whether two particles move in the same or opposite directions, while the latter includes more complex correlations such as twist and mutually perpendicular displacements.

3.2 Theory

We consider the dynamics of a protein in equilibrium in a given environment at temperature T. Assume that a trajectory in phase space is available from a reliable MD simulation. We first eliminate the overall translational and rotational motion because these are irrelevant for the internal motion we wish to analyze. The precise method of eliminating the overall motion is not important: either the linear and angular moments are removed every step in the simulation, or the molecular axes are constructed each step by a least-squares translational and rotational fit. The result in any case is a Cartesian molecular coordinate system in which the atomic motions can be expressed. The internal motion is now described by a trajectory $\boldsymbol{x}(t)$, where $\boldsymbol{x}$ can represent a subset of atoms. The correlation between atomic motions can be expressed in the covariance matrix C of the positional deviations:

$$C = cov(\boldsymbol{x}) = \langle (\boldsymbol{x} - \langle \boldsymbol{x} \rangle)(\boldsymbol{x} - \langle \boldsymbol{x} \rangle)^T \rangle \tag{3.1}$$

where $\langle\rangle$ denote an average over time. The symmetric matrix C can always be diagonalyzed by an orthogonal coordinate transformation T:

$$\boldsymbol{x} - \langle \boldsymbol{x} \rangle = T\boldsymbol{q} \text{ or } \boldsymbol{q} = T^T(\boldsymbol{x} - \langle \boldsymbol{x} \rangle) \tag{3.2}$$

which transforms C into a diagonal matrix $\Lambda = \langle \boldsymbol{q}\boldsymbol{q}^T \rangle$ of eigenvalues λ_i:

$$C = T\Lambda T^T \text{ or } \Lambda = T^T C T \tag{3.3}$$

The i^{th} column of T is the eigenvector belonging to λ_i. When a sufficient number of independent configurations (at least $3N + 1$) is available to evaluate C, there will be $3N$ eigenvalues, of which at least 6 representing overall translation and rotation are nearly zero. When a number of configurations, S, less then $3N+1$, is analyzed, the total number of nonzero eigenvalues is at most $S-1$ since the covariance matrix will not have full rank (see appendix B).

The matrix C has the property of being always connected to the holonomic constraints of the system. In appendix A we show that a subspace which is forbidden (or almost forbidden) for the motion is always fully defined by a subset of eigenvectors of the matrix C with zero (or approximately zero) eigenvalues. It is also important to note that the probability distribution of the displacements along the eigenvectors, although linearly uncorrelated, is not necessarily statistically independent. On the other hand, if a linear orthogonal transformation defines a subset of statistically independent generalized coordinates, then the unit vectors corresponding to this subset will always be eigenvectors of the covariance matrix C. The total positional fluctuation $\sum_i \langle (x_i - \langle x_i \rangle)^2 \rangle$ can be thought to be built up from the contributions of the eigenvectors:

$$\begin{array}{c} \sum_i \langle (x_i - \langle x_i \rangle)^2 > = \langle (\boldsymbol{x} - \langle \boldsymbol{x} \rangle)^T (\boldsymbol{x} - \langle \boldsymbol{x} \rangle) \rangle = \\ \langle \boldsymbol{q}^T T^T T \boldsymbol{q} \rangle = \langle \boldsymbol{q}^T \boldsymbol{q} \rangle = \sum_i \langle q_i^2 \rangle = \sum_i \lambda_i \end{array} \tag{3.4}$$

We choose to sort λ_i in order of decreasing value. Thus the first eigenvectors represent the largest positional deviations, and most of the positional fluctuations reside in a limited subset of the first n eigenvalues, where n is small compared to a total of $3N$.

3.3 Methods

Analysis was performed on the trajectories of two distinct simulations of hen eggwhite lysozyme.
A simulation in vacuum was performed by the authors, using the GROMOS simulation package and the GROMOS force field [15]. A starting structure was taken from the Brookhaven Protein Data Bank [36], entry 3LYZ. Including polar hydrogens, the system contained 1258 atoms. Nonpolar hydrogens were incorporated implicitly by the use of united atoms. In total a simulation of 1 ns was performed, with a step size of 2 fs. The temperature was kept at 298 K by coupling to an external temperature bath [37] , with a coupling constant $\tau = 0.01$ ps. Bond lenghts were constrained using the procedure SHAKE [16]. Rotational motion around, and translational motion of the center of mass was removed every 0.5 ps to prevent conversion of thermal motions into overall rotational and translational ones. Nonbonded interactions were evaluated using a short cutoff range of 0.8 nm, within which interactions were calculated every time step. Interactions in the range of 0.8-1.2 nm were updated every 20 fs. During the simulation, configurations were saved every 0.5 ps.

A. Mark kindly offered a 900 ps (100-1000 ps) trajectory of a simulation of lysozyme. This simulation, which included 5,345 water molecules, was performed at 300 K, also using the GROMOS package and and the corresponding force field [15] . Here configurations were saved every 0.05 ps. For further details concerning this calculation we refer to Smith et al [38].

Before the covariance matrix was built, all configurations were fitted to the first configuration by first fitting the center of mass and next performing a least square fit procedure [39] on the C_α coordinates. Covariance matrices C were constructed from the position coordinates of the atoms (all atoms or C_α atoms only) according to:

$$C_{ij} = \frac{1}{S} \sum_t \{x_i(t) - \langle x_i \rangle\} \{x_j(t) - \langle x_j \rangle\}, \tag{3.5}$$

where S is the total number of configurations, $t = 1, 2, ...S$, $x_i(t)$ are the position coordinates with $i = 1, 2, ...3N$, N is the number of atoms from which C is constructed and $\langle x_i \rangle$ is the average of coordinate i over all configurations. The system contained 129 C_α atoms, having 387 position coordinates. Eigenvalues and their corresponding eigenvectors were calculated using the **QL** algorithm [40]. Diagonalization of the C_α matrices, of size 387 by 387, required 24 s of CPU time on a single processor of a CONVEX 240, while diagonalizing the all atom covariance matrix of the solvent simulation of size 3792 by 3792 required 20.4 hr on the same machine.

3.4 Results and Discussion

Three different covariance matrices were diagonalized. The corresponding eigenvalues are shown in fig. 3.1, plotted in descending order against the corresponding eigenvector indices. Fig 3.1a shows the eigenvalues from the matrix that was constructed from (387) C_α coordinates in the vacuum simulation. In fig. 3.1b we show the eigenvalues as obtained from the (387) C_α coordinates from the solvent simulation. Finally, fig. 3.1c shows the eigenvalues obtained by analyzing the covariance matrix constructed from all atom coordinates (3792) of the protein from the solvent simulation. In this case the first few eigenvalues (mean square displacements) are one order of magnitude larger then in the previous plots, because the number of atoms involved in these displacements is approximately 10 times larger. Since the eigenvalues are mean square displacements, it is clear from fig. 3.1 that the configurational space of the protein is not a homogeneous space, in terms of the motion along the eigenvector directions. As can be seen from fig. 3.1a and b, the eigenvalues from the solvent simulation show a steeper decrease than those from the vacuum simulation. One reason for this may be the fact that the force fields used are not equivalent. In the vacuum force field, full charges have been replaced by dipoles. This produces a weakening of the electrostatic in-

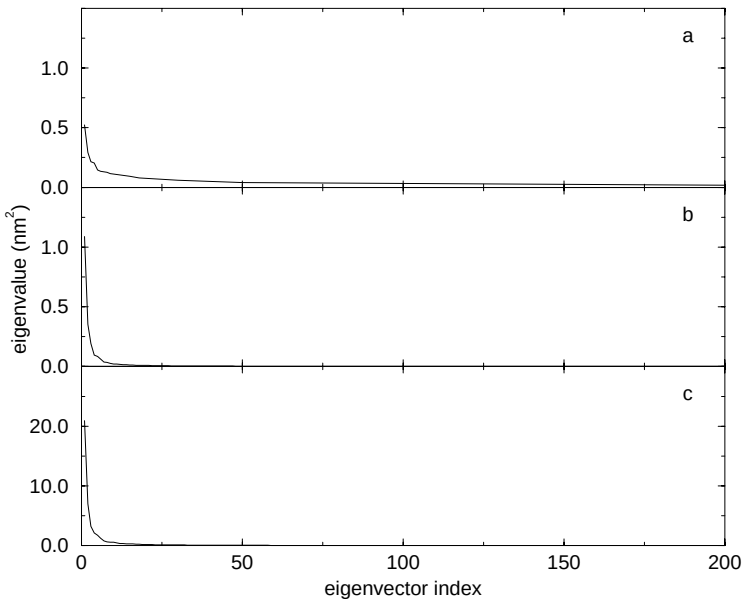

Figure 3.1: Eigenvalues, in decreasing order of magnitude, obtained from (**a**) C_α coordinates covariance matrix from the vacuum simulation; (**b**) C_α coordinates covariance matrix from the solvent simulation; (**c**) all atom coordinates covariance matrix from the solvent simulation.

teractions that, as we found, mainly affects the near constraints. As far as the methodology presented here is concerned there are no basic differences between vacuum and solvent simulation; so in the subsequent text we will show only the results obtained from the solvent simulation. At the end of this section the vacuum and solvent results will be compared.

The amount of motion associated to a subspace spanned by the first n eigenvectors can be defined as the corresponding subspace positional fluctuation (eq. 3.4, for the summation over n) where the eigenvalues are ordered in descending order. In fig. 3.2 we show this relative subspace positional fluctuation (with respect to the total positional fluctuation) versus the increasing number of eigenvectors that span the subspace. In

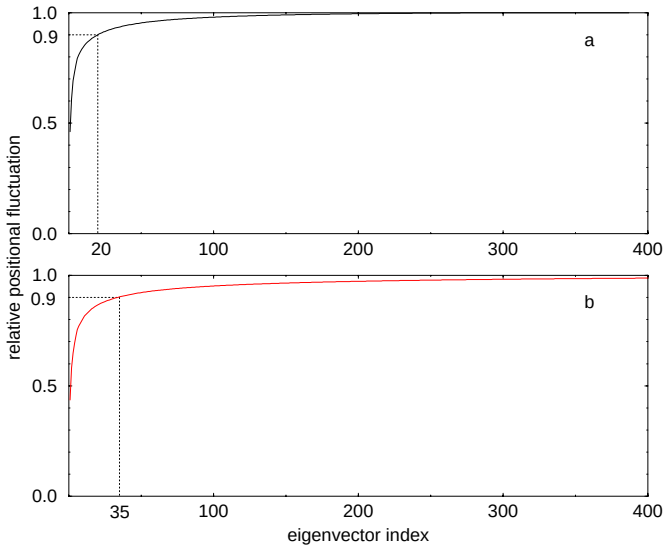

Figure 3.2: (**a**) Relative positional fluctuation (see text) of the motions along the eigenvectors obtained from the C_α coordinates matrix matrix (solvent simulation). (**b**) Relative positional fluctuation of motions along the eigenvectors obtained from the all atom coordinates covariance matrix (solvent simulation).

fig. 3.2a we show the results as obtained from the C_α matrix eigenvectors. Fig. 3.2b shows the results from the all atom analysis. From fig. 3.2 it can be seen that 90% of the total motion is described by the first 20 eigenvectors out of 387. If we analyze the motion due to all atoms (fig. 3.2b) we see that the first 35 eigenvectors out of 3792 contribute to 90% of the overall motion. This shows that most of the internal motion of the protein is confined within a subspace of very small dimension.

To have a closer look at the motion along the eigenvector directions one can project the trajectory onto these individual eigenvectors. In fig. 3.3 some projections of the C_α trajectory on the eigenvectors obtained from the C_α covariance matrix are plotted against time. It is clear from

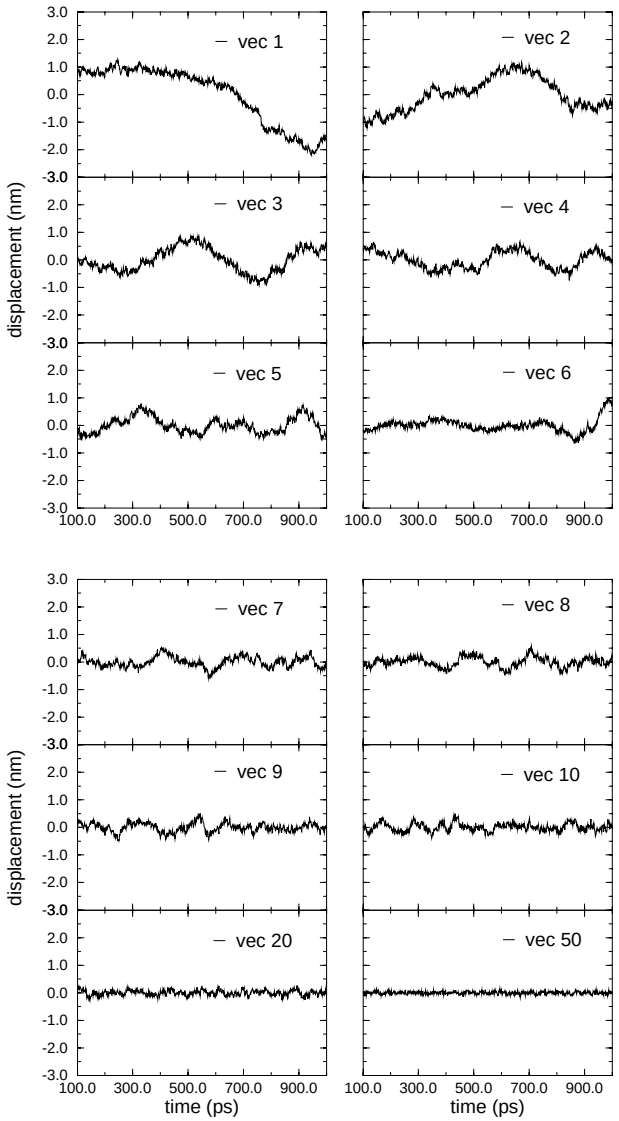

Figure 3.3: Motions along several eigenvectors obtained from the C_α coordinates covariance matrix (solvent simulation)

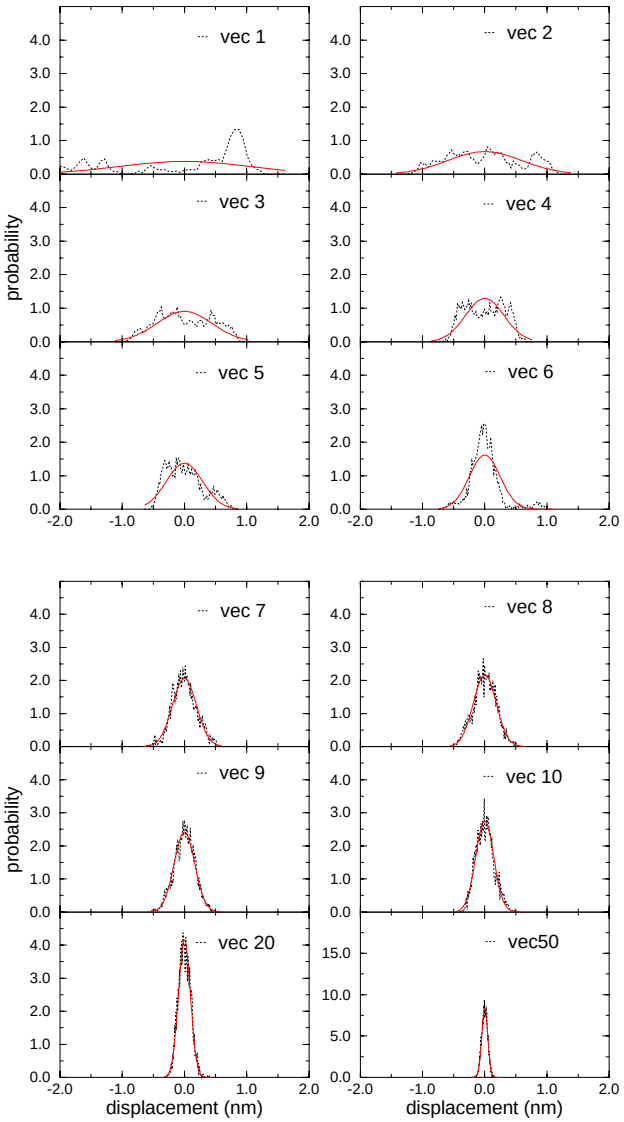

Figure 3.4: Probability distributions for the displacements along several eigenvectors obtained from the C_α coordinates covariance matrix (solvent simulation). Solid line: Gaussian distributions derived from the eigenvalues of the corresponding eigenvectors. Dashed line: sampling distributions.

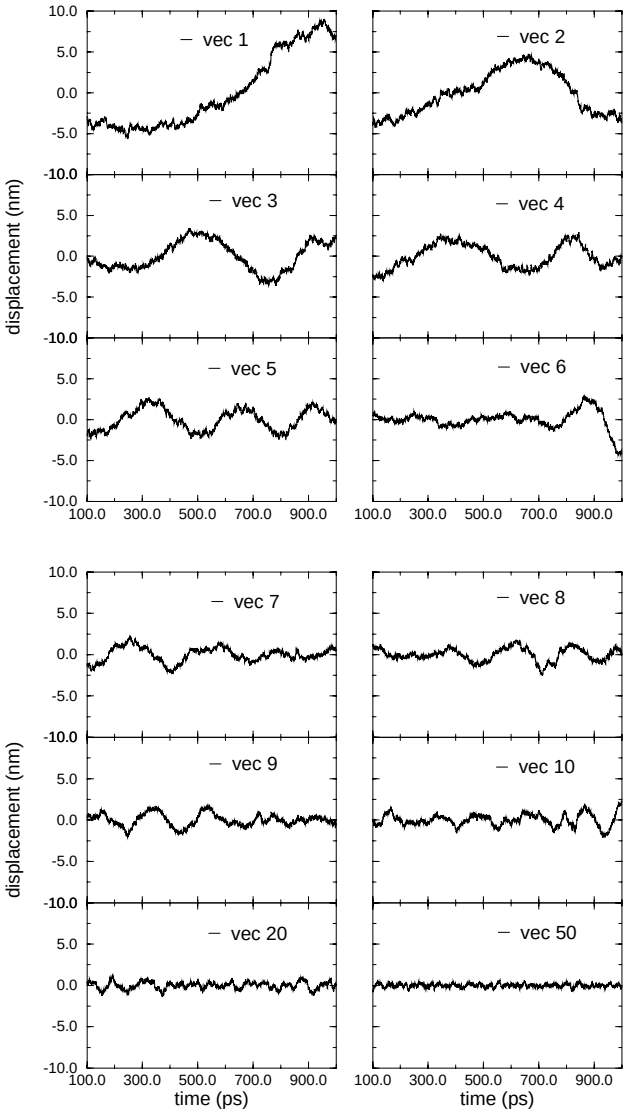

Figure 3.5: Motions along several eigenvectors obtained from the all atom coordinates covariance matrix (solvent simulation)

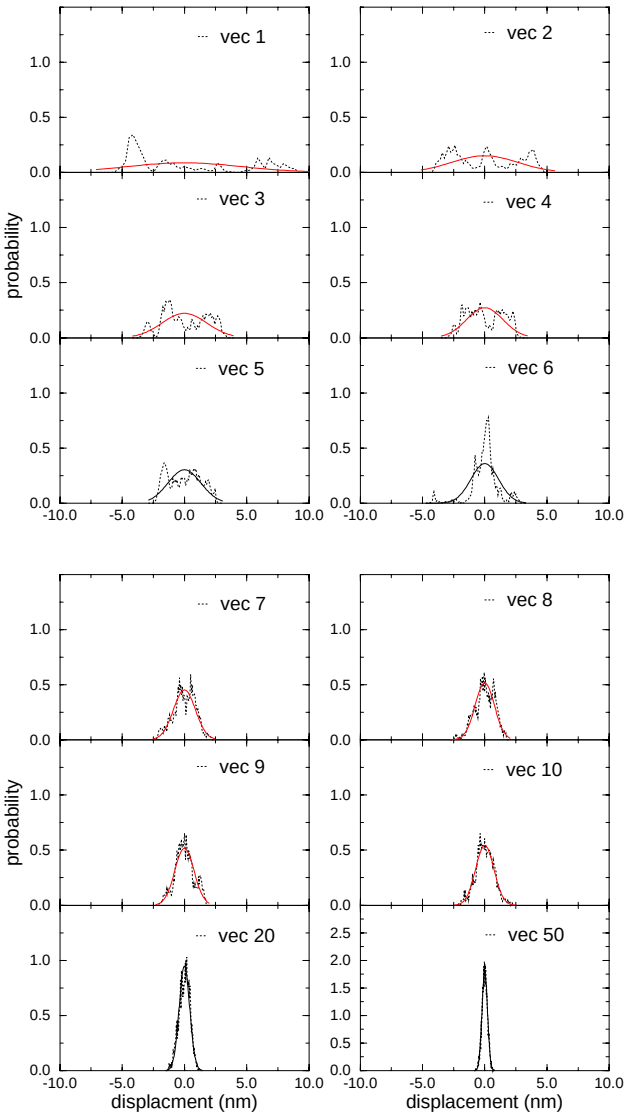

Figure 3.6: Probability distributions for the displacements along several eigenvectors obtained from the all atom coordinates covariance matrix (solvent simulation). Solid line: Gaussian distributions derived from the eigenvalues of the corresponding eigenvectors. Dashed line: sampling distributions.

this figure that all motions that have not yet reached their equilibrium fluctuation belong to the first 10 eigenvectors. Fig. 3.4 shows the observed distribution functions for the displacements along the same eigenvectors, as well as the corresponding Gaussian functions with the same variance and average value. Obviously the only non-Gaussian distributions are again found within the first 10 eigenvectors. Fig. 3.5 and 3.6 show the same, but now projections and distributions have been evaluated using the eigenvectors that were obtained from the all atom covariance matrix. Just as in the case for the C_α analysis we find that all the motions that have not yet reached equilibrium fluctuation are confined within the first 10 eigenvectors. Also the only non-Gaussian distributions appear within the same eigenvectors.

We also noticed a great similarity between the motions along the first few eigenvectors of the C_α matrix and those along the first few eigenvectors derived from the all atom matrix. To investigate this similarity further, we extracted the components from the all atom eigenvectors that corresponded to the C_α coordinates and normalized the vectors that we obtained in this way. In fig. 3.7 the projections of these vectors on the eigenvectors of the C_α matrix are plotted. It is clear that the first 8 extracted vectors correspond to the first 8 C_α matrix eigenvectors. It should be mentioned that the length of these extracted vectors is approximately 20% of the whole length of the corresponding all atom eigenvector, whereas the total number of C_α atoms is about 10% of the total number of atoms in the protein. This indicates that the essential internal motion of the protein mainly involves the backbone atoms. We also noted that the displacements along the first 5 eigenvectors produced a large motion near the active site of the molecule. Fig. 3.8 shows a superposition of 10 sequential projections of the C_α motion onto the first eigenvector, each separated by 100 ps (compare with fig. 3.3). The catalytic site residues Glu-35 and Asp-52 are rigid, but the entrance to the active site cleft, including residues involved in substrate binding (59,62,63,101,107) [41] shows extensive flexibility. This motion, which also involves other loops in the protein, possibly affects the association and dissociation of sub-

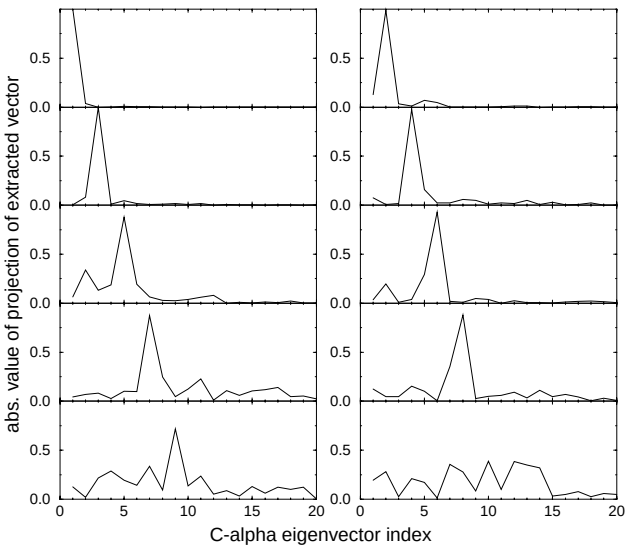

Figure 3.7: Absolute value of the projections of the (normalized) extracted vectors coming from the first 10 eigenvectors of the all atom coordinate covariance matrix (see text) on the eigenvectors obtained from the C_α covariance matrix (solvent simulation)

strates and products.

Fig. 3.9 shows the trajectory projected on four planes, each defined by two all atom matrix eigenvectors. In the planes of fig 3.9a and b (respectively, eigenvectors 1 and 2 and eigenvectors 2 and 3) the trajectories are confined within narrower ranges than those expected from independent motions, suggesting the presence of a coupled force field. In fig. 3.9d (eigenvectors 20 and 50) the trajectories fill the expected ranges almost completely. This means that we are dealing with basically independent motions. We analyzed the vacuum simulation and compared the motion in the essential space with that of the solvent simulation. The motion in the vacuum simulation appears to be largely restricted to the carboxy terminal strand; the motion near the active site is no longer present.

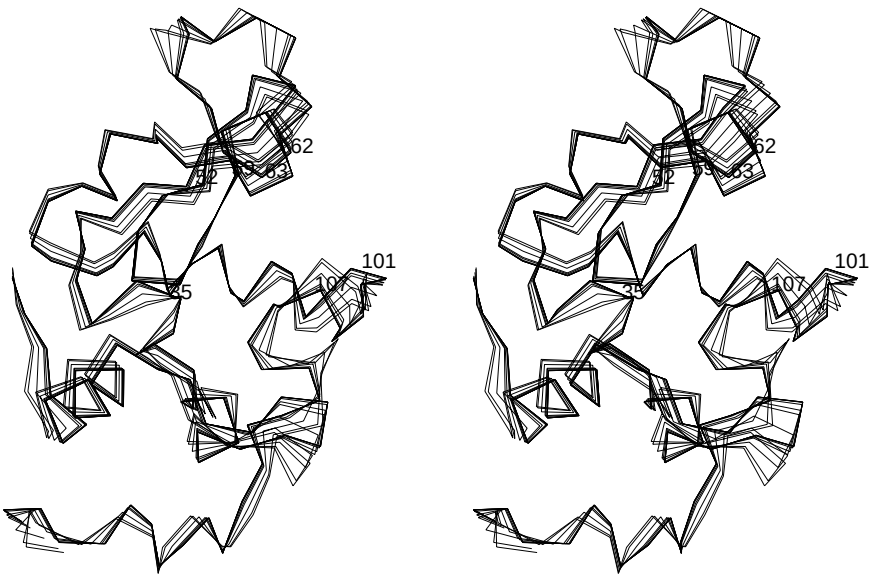

Figure 3.8: Superposition of 10 configurations obtained by projecting the C_α motion onto the first eigenvector. Configurations are separated by 100 ps. Residues involved in the catalytic reaction (35 and 52) and in the binding of the substrate (59,62,63,101, and 107) are indicated

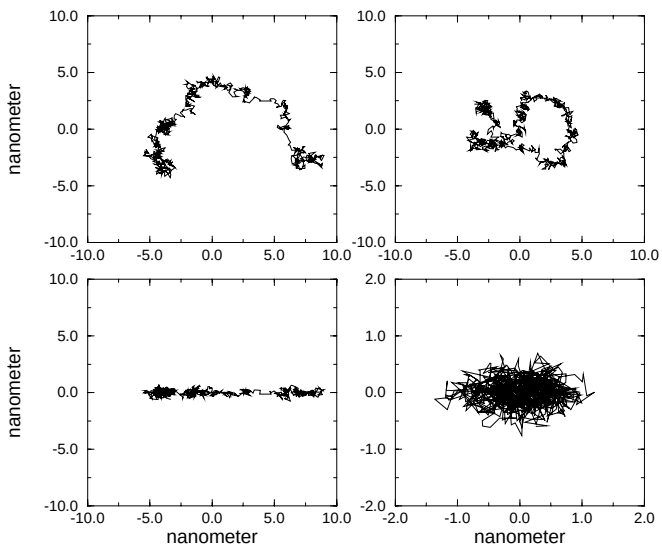

Figure 3.9: Projection of the trajectory (solvent simulation) on the planes defined by two eigenvectors from the all atom coordinates covariance matrix. (**a**) Horizontal axis: displacement along first eigenvector. Vertical axis: displacement along second eigenvector. (**b**) Horizontal axis: displacement along second eigenvector. Vertical axis: displacement along third eigenvector. (**c**) Horizontal axis: displacement along first eigenvector. Vertical axis: displacement along 50^{th} eigenvector. (**d**) Horizontal axis: displacement along 20^{th} eigenvector. Vertical axis: displacement along 50^{th} eigenvector. (note the difference in scale)

3.5 Conclusions

The analysis given in this article shows that the essential dynamics of lysozyme, and presumably of other globular proteins, can be described in a subspace of very small dimension (less then 1% of the original Cartesian space) consisting of linear combinations of Cartesian degrees of freedom defined in a molecule-fixed coordinate system. All other degrees of freedom can be considered as corresponding to irrelevant Gaussian fluctuations, behaving like near-constraints. The essential subspace itself is defined by the near-constraints, which are related to the mechanical structure of the molecule in a given conformation. We have strong evidence from inspection of a few proteins studied up to now (lysozyme, thermolysin, and a subtilisin analog) that these motions are related to the functional behavior of the proteins such as opening and closing of the active site and hinge-bending motions between two domains enclosing the active site. The analysis of this behaviour will be the subject of a subsequent study. A (major) conformational change to a different folded conformation may alter the characteristics of the essential subspace, while unfolding will lead to an increase of its dimensionality. The fact that active site motions are not present in the essential space of the vacuum simulation suggests strongly that vacuum simulations are not suitable for the study of biologically relevant motions.

3.6 Acknowledgments

We thank Dr. A.E. Mark from ETH in Zürich for offering a 900 ps trajectory of lysozyme in solvent. We are also grateful to Daan van Aalten who applied the method presented in this paper to a few other proteins. Finally, we gratefully acknowledge the Italian foundation "Cenci Bolognetti", Istituto Pasteur, for financially supporting Dr. A. Amadei in this project.

Chapter 4

An extended MD simulation of wild-type haloalkane dehalogenase

4.1 Introduction

For wild-type dehalogenase, the rate-limiting step in the overall kinetics of dehalogenation of 1,2-dichloroethane is the release of the halide ion [14], a product of formation of the intermediate ester. To find out whether the presence of a halide ion, in the active site cavity, has any influence on the dynamics of the protein, two MD simulations were performed. In one of these simulations a chloride ion was bound in the active site, and this calculation will be the subject of the next chapter. This chapter deals with the simulation of dehalogenase in the absence of halide. We will have a closer look at the stability of the molecule and investigate the types of motion that occurred during the simulation.

4.2 Methods

Simulations were performed using the GROMOS force field [15] with adjusted Lennard-Jones interactions between aliphatic carbon atoms and water oxygen atoms [42] and explicit hydrogens added to aromatic residues [43]. Polar hydrogens were included explicitly whereas nonpolar hydrogens were implicitly included by the use of united atoms. As a starting structure, the X-ray structure at pH 8.2 and a resolution of 1.9 Å (Brookhaven data bank entry 1EDE) [3, 7, 44] was chosen. All histidines were assumed to be electrically neutral. To find out which ring nitrogen was protonated, the X-ray structure was checked for the presence of possible hydrogen bonds by looking at the closest hydrogen bond acceptor. Histidine residues 37, 54 and 289 were protonated at $N_{\delta 1}$ positions, residues 102 and 305 at the $N_{\epsilon 2}$ positions.
Periodic boundary conditions were applied by the use a truncated octahedron filled with equilibrated water molecules. Each water molecule of which the oxygen atom had a distance to any non-hydrogen protein atom (or water molecule present in the X-ray structure) of less then .23 nm was removed. The minimum distance between protein atoms and the walls of the box was taken as 0.75 nm. The volume of the box was 176 nm^3. The protein possessed a negative charge of -17 e. To obtain an electrically neutral system, 17 sodium ions were added as counterions. This was done by calculating the electric potential at the positions of all oxygen atoms of water molecules and replacing the 17 water molecules with the lowest potential by a sodium ion. In total the system contained 16287 atoms of which 3169 were part of the protein. Subsequently the energy of the system was minimized by the steepest descent method for 100 steps without position restraining.
During the first 10 ps of simulation, harmonic position restraining was applied to all protein atoms, using a force constant of 9000 kJ nm^{-1} mol^{-1}. After that, the position restraints were removed and the system was simulated for 1 ns. Initial velocities were taken from a Maxwell-Boltzmann distribution at 298 K. The temperature was kept constant by coupling solute and solvent separately to a thermal bath [37] at 298 K

with a coupling constant τ_T=0.1 ps. Pressure was kept constant by coupling to a pressure bath at 1.0 bar [37] using a coupling constant τ_P=0.5 ps. Bond lengths were constrained using the SHAKE method [16] with a relative tolerance of 0.0001. Nonbonded interactions were calculated using a twin cut-off radius: within a short cut-off radius of 0.8 nm interactions were calculated every timestep, all other interactions within a radius of 1.2 nm were only calculated every 10 steps. The atom pair list for the short range interactions was updated every 10 steps. The stepsize was 0.002 ps.
Simulations were performed using the GROMOS87 software package [15]. Nonbonded routines were rewritten by the author to be used on a Cray J90 parallel computer. The simulation took ∼200 hrs of real time. Analysis of secondary structure elements, rmsd, B-factors and radius of gyration was performed with GROMACS [45].

4.3 Results and discussion

To investigate the stability of the enzyme, three properties were investigated: the stability of secondary structure elements, the rmsd of the protein with respect to the crystal structure and the radius of gyration of the protein. Fig. 4.1 shows the behaviour of the secondary structure elements, as calculated with the program DSSP [46]. For clarity, the elements are subdivided into only three catagories: α-helices, β-sheets and and coils (including β-bridges, bends, turns, π-helices and 3_{10}-helices). Although there are fluctuations, in general the secondary elements remain intact. One exception is helix 5 (residues 171-181), where the helicity is lost after Trp175. The radius of gyration R_g as a function of time is shown in fig 4.2. If unfolding occurs then this is often revealed by an increase of R_g. Initially R_g increases slightly, but stabilizes after 600 ps, resulting in a deviation of less then 2 % from the X-ray structure. The root mean square deviation (rmsd), of the backbone C_α atoms, from the X-ray structure is shown in fig 4.3. Here we notice that a relatively large

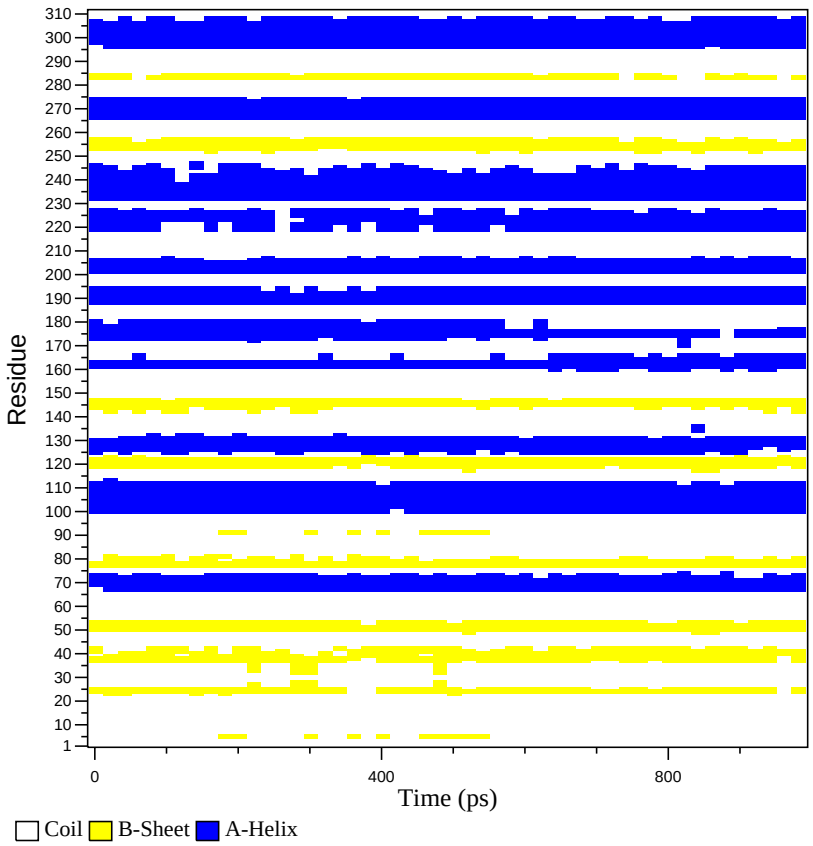

Figure 4.1: Secondary structure elements vs. time

rmsd arises, but again there appears to be a stabilization after 600 ps. To find out which regions are involved in this large rmsd, B-factors kan be derived from the trajectory. B-factors are defined as:

$$B = \frac{8}{3}\pi^2 \langle |\Delta \boldsymbol{r}|^2 \rangle \tag{4.1}$$

where $\langle |\Delta \boldsymbol{r}|^2 \rangle$ is the mean square atomic displacement. Fig 4.4 shows the B-factors from the simulation (for C_α atoms only) compared to the crystallographic ones. For several regions, the fluctations are one order of

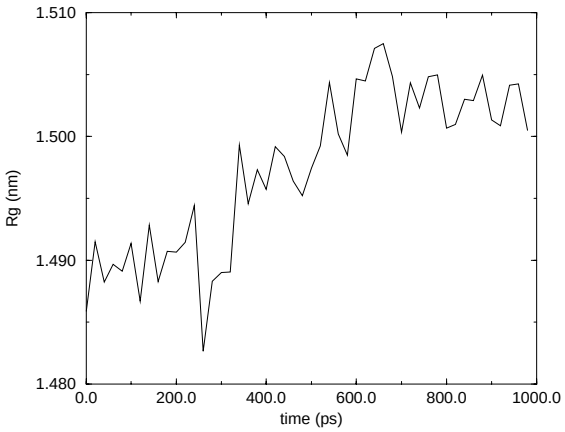

Figure 4.2: Radius of gyration vs. time

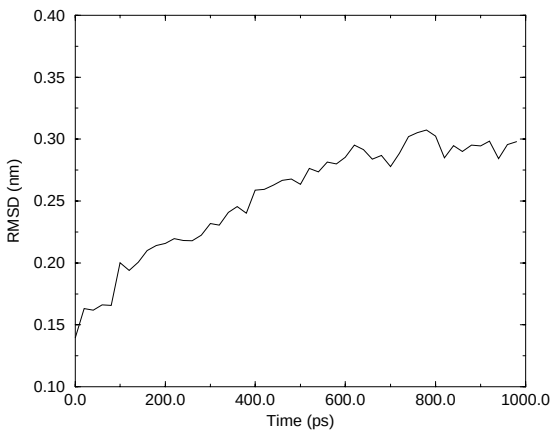

Figure 4.3: Root mean square deviation from the crystal structure vs. time

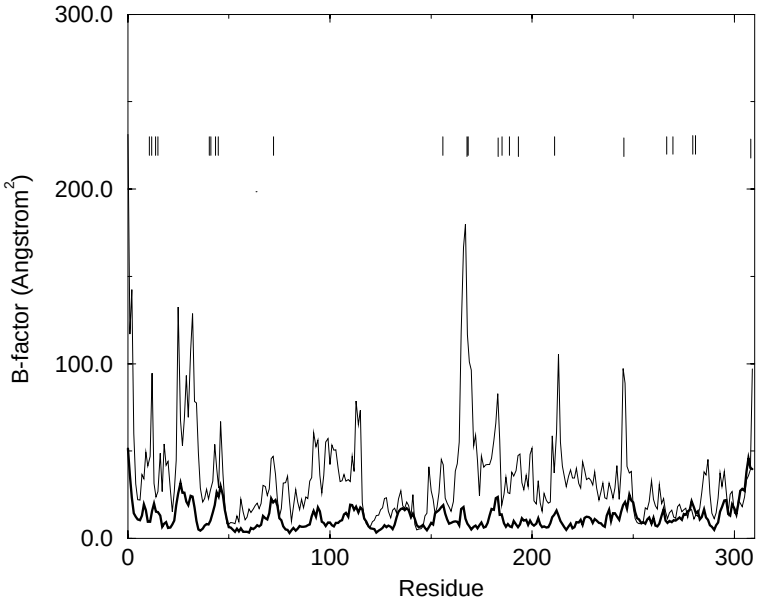

Figure 4.4: B-factors, thick line: crystallographic B-factors, thin line: B-factors derived van the simulated trajectory. The short thin lines indicate crystallographic contacts. The highest peak is positioned at Pro168.

magnitude larger in the simulation compared to the X-ray structure. But peaks in the simulation always correspond to peaks in the X-ray structure. Fig 4.4 also shows that the region where the largest fluctuations occur, the cap domain (residues 156-229), is also rich in crystallographic contacts, especially close to peaks.

The trajectory was further investigated by applying Essential Dynamics (ED) analysis [6] described in section 2.3.1. In our case we constructed a covariance matrix from the 930 C_α coordinates in the time interval 100-1000 ps. The ten largest eigenvalues from this matrix are shown in fig 4.5 (in decreasing order of magnitude). The projections along the first six eigenvectors are shown in fig 4.6. One might look at

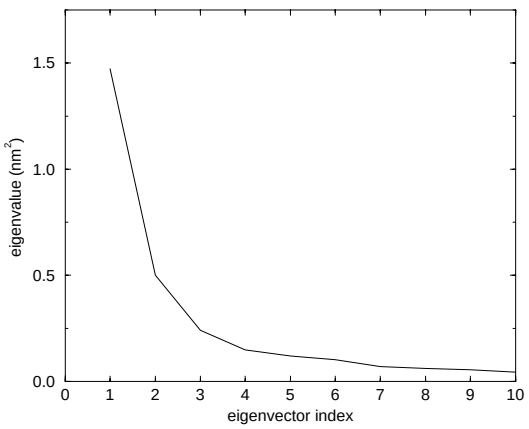

Figure 4.5: The 10 largest eigenvalues in order of decreasing magnitude

the motions produced by projection of the trajectory on separate eigenvectors. It has been suggested however [47] that the motion within the essential subspace is of a diffusive nature. We found further support for this by analysing the trajectory of a multidimensional Brownian particle (appendix C). For this reason it is not obvious that the eigenvector with the largest eigenvalue, which corresponds to the direction in which the largest fluctuation happens to occur, also corresponds to a direction in which the largest fluctuation is possible. So we chose to look at the motion, produced by projection of the trajectory on the hyperplane defined by the first three eigenvectors. The total mean square displacement (msd) of the C_α carbon atoms was 3.78 nm^2, whereas the sum of the first three eigenvalues was 2.21 nm^2. This means that in this way 58% of the total msd was retained. Fig. 4.7 shows the B-factors for the C_α atoms, as produced by the first three eigenvectors as well as the B-factors produced by the total motion along all other directions. Especially the motion in the cap-domain (residues 156-229) is mainly produced by the first three eigenvectors. It can also be seen that the residues, involved in catalysis (Asp124, Trp125, Trp175, His289 and Asp260) all have little

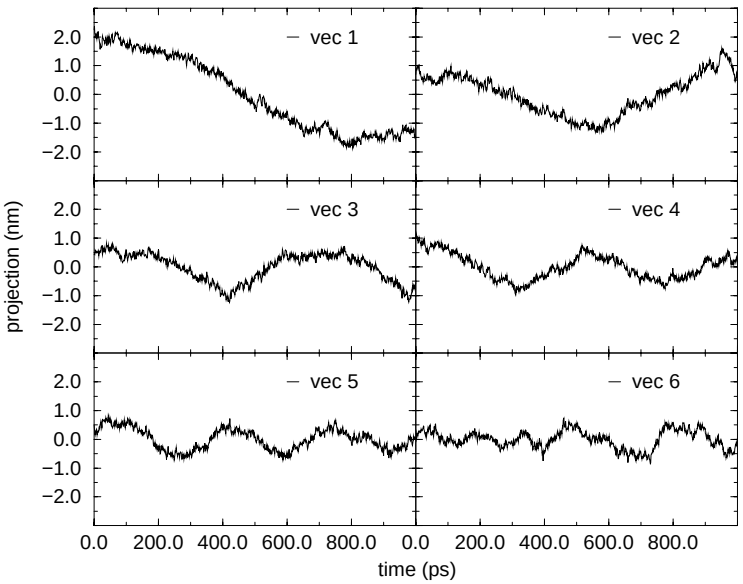

Figure 4.6: projections along the first 6 eigenvectors

motion except for Trp175. But even here the B-factor lies in a minimum with respect to its direct environment. There is a small peak at Asp260, probably caused by the fact that the hydrogen bond between this residue and His289 is lost. Fig. 4.8 shows the motions along the first three eigenvectors as they appear in three dimensional physical space.

After this we performed rigid body analysis (section 2.4) on the trajectory. In principle, this can be done on the full trajectory, as well as on the trajectory obtained by projection on the first three eigenvectors. In fig 4.9 the scores for both, from the time interval 900-1000 ps are shown. There is no significant difference between them. To interprete these data we have to remember that the magnitude of the distance fluctuations between two Cα atoms is related to the absolute value of the difference between their rigidity scores. This means that a group of atoms having similar scores also has small interatomic distance fluctuations and can be considered as being rigid. Based on the rigidity scores, we can roughly

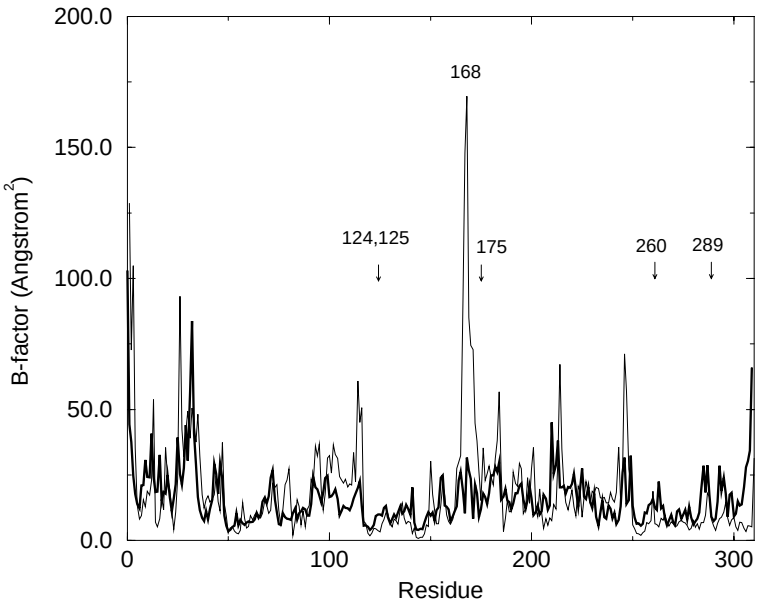

Figure 4.7: B-factors derived from the total motion in the hyperplane of the first three eigenvectors (thin line) and of the total motion in all other directions (thick line). Asp124, Trp125, Trp175 and His289 are catalytic residues. There is a sharp peak at Pro168.

divide the protein into three regions, A, B, and C, with respectively positive, neutral, and negative scores. The boundaries between them are indicated in the figure. The distance fluctutations between regions A and C are largest. In fig 4.10 the regions are visualized in the three-dimensional structure. Apparently, segment 183-200 should be considered as a separate region. So we subdivide A into A_I (residues 183-200) and A_{II} (residues 1-3,24-37 and 100-114). The segment between residues 149 and 182, assigned to region C, has a rather irregular pattern, meaning that there are also considerable internal distance fluctuations present. A strong negative peak is observed at Pro168 indicating that this residue

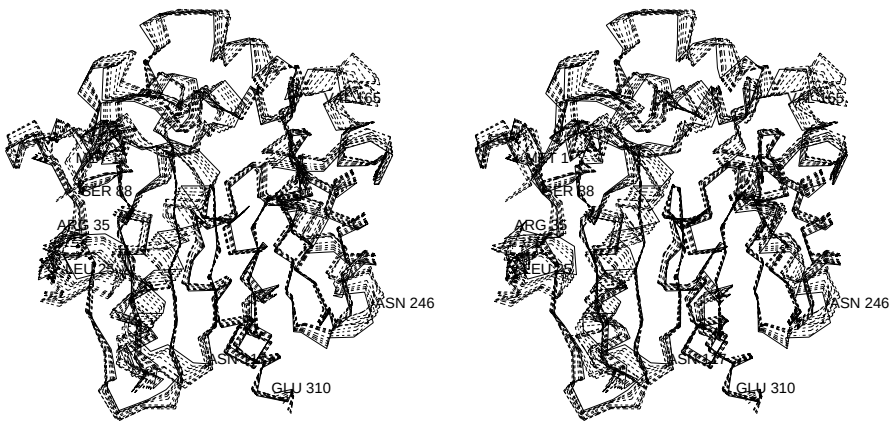

Figure 4.8: 11 structures taken from the projection of the trajectory on the first three eigenvectors; thick solid line: projection of the starting structure, thick dashed line: projection at 1ns, thin dashed lines: nine projections seperated by 100ps. (drawn with WHATIF [48]).

is rather flexible with respect to any other region in the protein. Trp175 (involved in halide ion binding), has a score that is most similar to the one of region B, which covers most of the main domain, and also contains other residues that are involved in catalysis. We also looked for a possible tunnel, connecting the active site cavity with the exterior and serving as an entrance for the substrate. In the crystal structure such a tunnel was found [7] but it was blocked by the sidechain of Leu262. The solvent accessible surface of the structure, after 1 ns, was calculated with WHATIF [48], and the result was visually inspected on the presence of a connection of the active site with the surface of the protein. No such opening was found.

4.4 Conclusions

During the simulation the protein appeared to possess a large flexibility, giving rise to large rms deviations from the crystal structure that stabilized after $\sim$ 600ps. Based on atomic distance fluctuations, the molecule can roughly be subdivided into four regions which were designated as A_I, A_{II}, B and C. The term *rigid body* does however not apply to segment 149-183, as this part of the molecule appears to possess a relatively large internal flexibility. It is interesting to note that Trp175, which is part of this region, has only small distance fluctuations with respect to all other catalytic residues that are found in region B which covers most of the main domain. The fact that also the B-factors derived from the simulation are small for all the catalytic residues may further indicate that, in spite of the large rms for the complete structure, the protein is not unfolding.

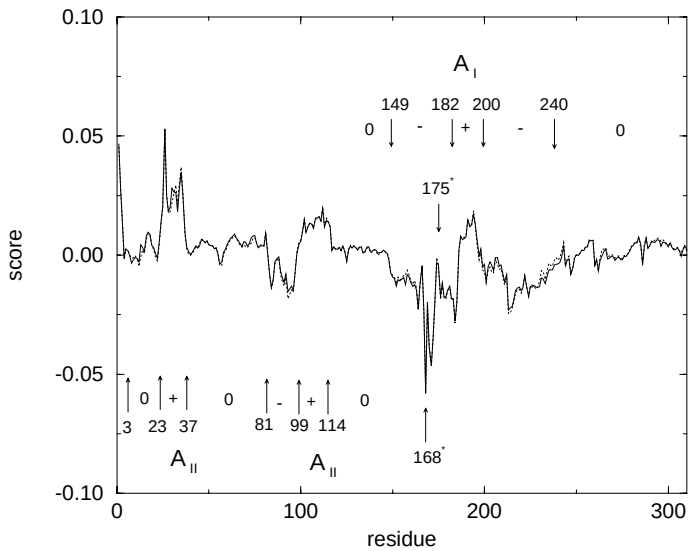

Figure 4.9: Rigidity scores, derived from the time interval 900-1000 ps of the full trajectory (solid line) and the projection on the first three eigenvectors (dotted line). 0, + and - indicate neutral, positive or negative scores. Trp175 lies in the middle of a region with negative scores, but has itself a neutral score. Pro168 shows irregular behaviour and has a strong negative peak. Based on visual inspection, region A is subdivided into A_I and A_{II}

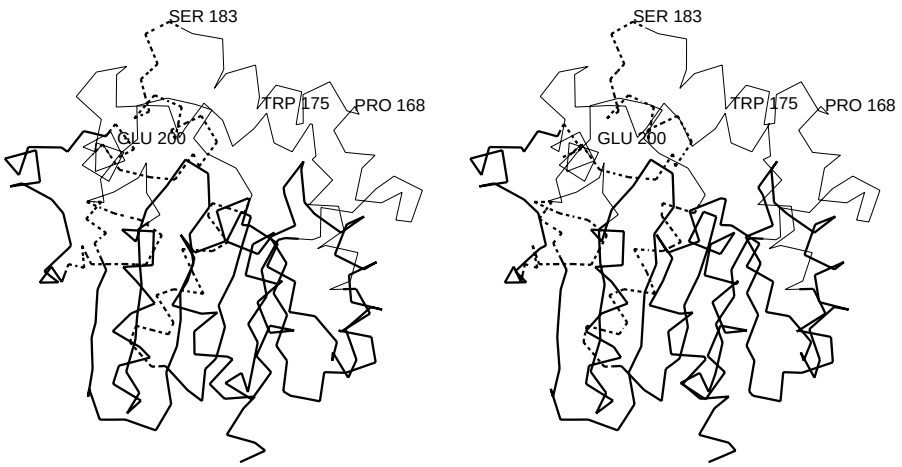

Figure 4.10: Regions A (thick dashed line), B (thick solid line) and C (thin solid line), obtained by subdividing the molecule based on rigid body scores. The largest distance fluctuations are found between regions A and C. Region A consists of two separate regions. Compared to regions A and B, region C has a relatively large internal flexibility: Trp175 has a neutral score (and should therefore be assigned to region B), whereas Pro168 has a strongly negative score and appears to be flexible relative to any other part of the protein.

Chapter 5

An extended MD simulation of wild-type haloalkane dehalogenase including a chloride ion bound in the active site

5.1 Introduction

To find out the effect of choride bound in the active site, simulations including the ion in the active site were performed. Initially, this was done using the usual GROMOS force-field [15], but the interaction of the ion with its hydrophobic environment, containing several aromatic residues, was too weak, so chloride almost instantaneously (within 20 ps) left the active site cleft. In this chapter we describe and analyse an MD simulation in which we have added a polarizability term to stabilize the bound state of the ion.

5.2 Methods

For the simulations the same software and force-field were applied as described in section 4.2, with a polarizability term added (section 2.2). As a starting structure, the X-ray structure at pH 6 and a resolution of 2.1 Å (Brookhaven data bank [36] entry 2DHE) [9, 7, 44] was taken. Except for His289, all histidines were assumed to be electrically neutral. To find out which ring nitrogen was protonated, the X-ray structure was checked for the presence of possible hydrogen bonds by looking for the closest hydrogen bond acceptor. Histidine residues 37 and 54 were protonated at $N_{\delta 1}$ positions, residues 102 and 305 at the $N_{\epsilon 2}$ positions. To compensate for the negative charge on the chloride ion, His289 was protonated at both positions making it positively charged. The volume of the periodic box (truncated octahedron) was 174 nm^3. The protein possessed a negative charge of -17 e. To obtain an electrically neutral system 17 sodium ions were added as counterions by evaluating the electrostatic potential at the oxygen positions of all water molecules, and replacing those 17 water molecules with the lowest potential by a sodium ion. In total, the system contained 16046 atoms of which 3171 were protein atoms (including the bound chloride ion). Only the chloride ion was treated as a polarizing atom (section 4.2). To decide which atoms should be polarizable, initially all residues that contained at least one atom within a distance of 5 Å (in the X-ray structure) from the chloride ion were taken and all atoms from these residues were treated as polarizable atoms. The residues found were: Glu56, Asp124, Trp125, Phe128, Phe172, Trp175, Phe222, Pro223 and Val226. However, Asp124 had only one atom within 5 Å ($O_{\delta 2}$ with a distance of 4.9 Å). One would like to have as few charged residues included as possible, given the fact that the polarization model was designed for describing interaction of ions with hydrophobic residues. For this reason, and because it's distance was such that polarization effects are small compared to Coulombic interactions, Asp124 was excluded. The energy of the system was minimized, without restraints, by the steepest descent method for 100 steps. After this, it was simulated for 10 ps using harmonic position restraints on all pro-

tein atoms, with a force constant of 9000 kJ nm^{-1} $mole^{-1}$. Then, the position restraints were removed and the system was simulated for 1 ns. Later it was continued for another 500 ps. Analysis of secondary structure elements, rmsd, B-factors and radius of gyration was performed with GROMACS [45].

5.3 Results and discussion

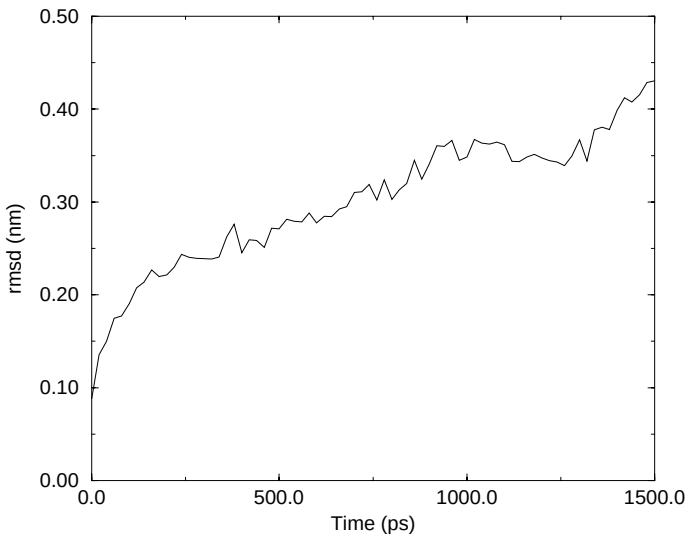

Figure 5.1: Root mean square deviation from the crystal structure vs. time.

As was already mentioned, originally a simulation of 1 ns was performed. But the rmsd showed a steady increase, which initially led us to believe that the protein was unfolding. To find out how the system would develop further the simulation was continued for another 500 ps. The resulting

rmsd as a function of time is shown in fig. 5.1. After 1 ns the rmsd stabilizes around a value of approximately 0.35 nm to increase again after ∼1.4 ns. Close inspection of the trajectory showed that at ∼1.25 ns some (charged) residues of the protein in the central periodic box started to interact with residues from the protein's images, i.e. the distance between them became smaller than the large cut-off radius (1.2 nm), and later even smaller than the short cut-off radius (0.8 nm). For this reason we only considered the trajectory until 1.2 ns as being reliable, and restricted our further analysis to this time interval. Fig 5.2 shows the radius of

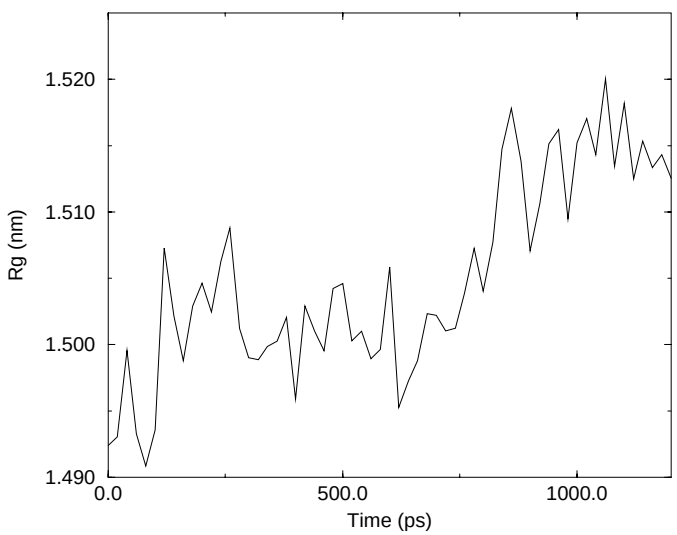

Figure 5.2: Radius of gyration vs. time.

gyration as a function of time. Here we see a substantial increase in the time interval between 600 and 900 ps, after which there seems to be a stabilization. In fig. 5.3 the crystallographic B-factors, as well as the B-factors derived from the simulation are plotted. It can be seen that most of the mean square displacement is concentrated in the region between residues 180 and 210 with distinct peaks at Ser183 and Thr197. There

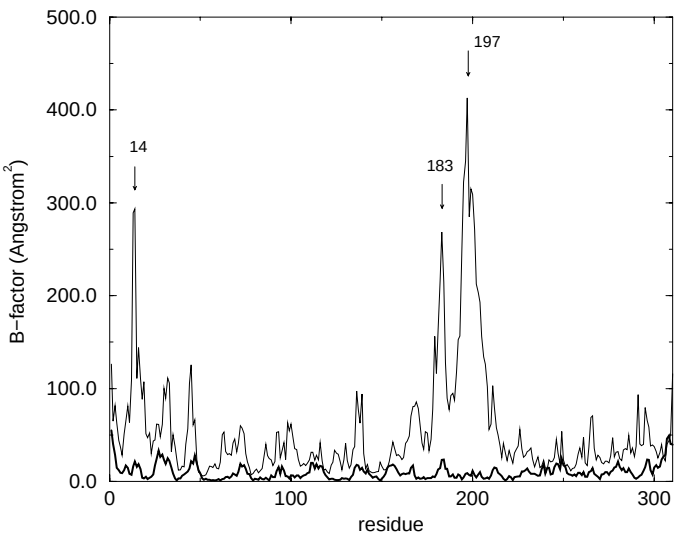

Figure 5.3: B-factors from C_α atoms, thick line: crystallographic B-factors, thin line: B-factors derived from the simulated trajectory. Peaks at Asn14, Ser183 and Thr197 are indicated.

is also a more narrow peak, with a maximum at Asn14. In fig. 5.4 the secondary structure elements versus time are plotted. Comparison with the previous simulation (fig. 4.1) shows some major differences. Helices 3,4 and 8 (residues 125-136,159-166 and 217-227) are less stable in the current simulation. But helix 5 (residues 171-181) appears to be more stable. One might expect that especially in the region between residues 180 and 210 secondary structure elements are unstable, as in this region the largest rmsd is found. Strangely enough however, both helices in this region are highly stable.

For Essential Dynamics (ED) analysis, a covariance matrix was constructed from the C_α coordinates, from the time interval 100-1200 ps. Fig 5.5 shows the 10 largest eigenvalues. As is to be expected from the large rmsd, they are considerably larger than the ones found in the

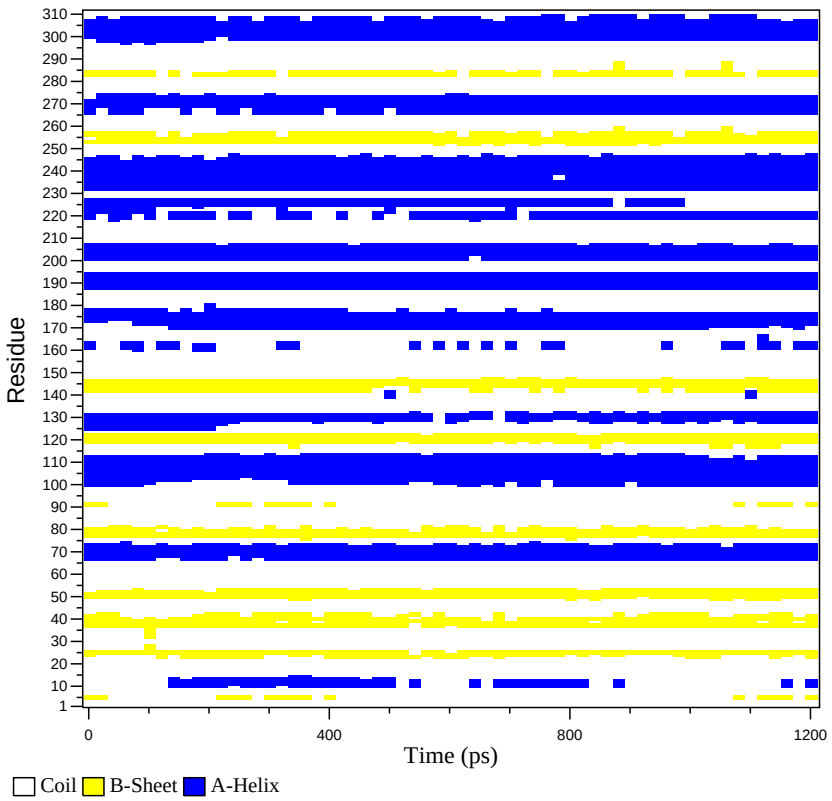

Figure 5.4: Secondary structure elements vs. time.

previous chapter (fig 4.5). The projections along the first 6 eigenvectors are shown in fig. 5.6. The motion along the first 3 eigenvectors in 3 dimensional physical space is shown in fig. 5.7. The total mean square displacement (msd) was 5.82 nm^2 and the sum of the first 3 eigenvalues was 4.01 nm^2. This means that fig. 5.7 shows 68 % of the total msd. It also shows that most of this motion takes place in the cap domain. Helix 5 (residues 171-181) shifts to the right. The motion of chain segment 180-210 seems to be rather disorganized. To obtain a clearer notion of the type of motion we used the same trajectory for rigid body analysis (section 2.4).

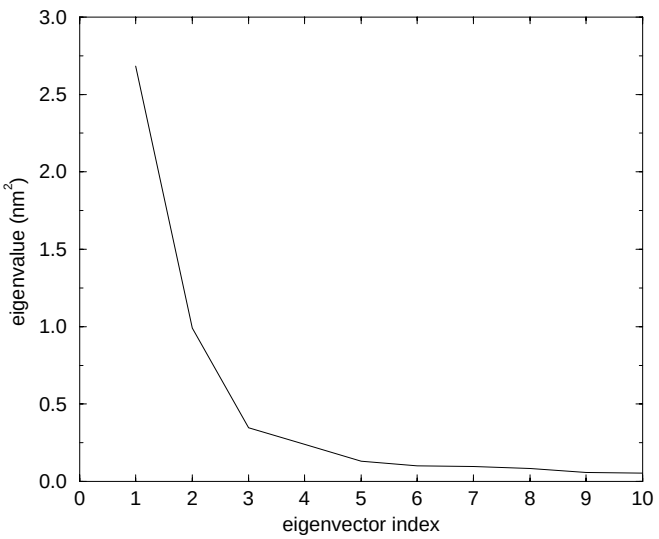

Figure 5.5: The 10 largest eigenvalue in order of decreasing magnitude.

The rigidity scores are shown in fig. 5.8. If we compare fig. 5.8 with the results from the previous chapter (fig. 4.9), we find large differences. The peak with a maximum at residue Thr197 immediately catches the eye. Also fig. 4.9 shows a peak, but not so outstanding. Between Cys150 and Val180, an irregular pattern was found in fig. 4.9. This has turned into a much more regular peak. The peaks in segments 1-3, 23-37, and 81-114 seem to have disappeared. Instead, there is now a region with positive scores with an irregular pattern between residues 253 and 295. If we now subdivide the molecule we find: region A containing residues 184-211, region B containing residues 1-149, 212-252 and 296-310, and region C containing residues 150-183 and 253-295. These regions do not correspond to the ones found in the previous chapter. In fig. 5.10 the new regions are visualized. The regularity of the peaks with maxima Pro168 and Thr197 suggests rigid bodies connected by a hinge. To verify this we performed a least square fit of both regions (skipping residues 181-

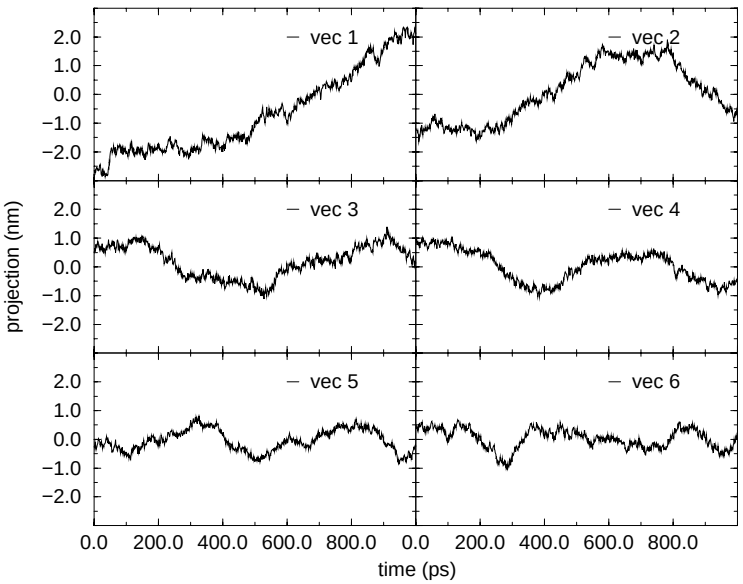

Figure 5.6: Projections along the first 6 eigenvectors.

184) from the MD structure after 1.2 ns on the corresponding regions of the crystal structure. The results are shown in fig. 5.9. Compared to the large motions in which both parts of the molecule are involved, they show only little internal motion and there is no clear sign of unfolding. The seemingly disorganized motion in fig. 5.7 is in fact an opening of a helix-loop-helix region between residue 185 and 211. This can be seen more clearly in fig. 5.13 and 5.14. We also had a closer look at the behaviour of the chloride ion. Fig. 5.11 shows the ion and its surroundings in the crystal structure. It is clearly bound between $N_{\delta 1}$ of Trp125 and Trp175. Fig. 5.12 shows the same but now after 1.2 ns of simulation. Drastic changes have taken place. Its interaction with Trp125 is weakened (the distance increases from 3.1 Å to 5.1 Å) whereas there are strong interactions with peptide nitrogens from Lys224, Met225 and Val226. Here it should be stressed that Lys224 and Met225 are not taken to be polarizable.

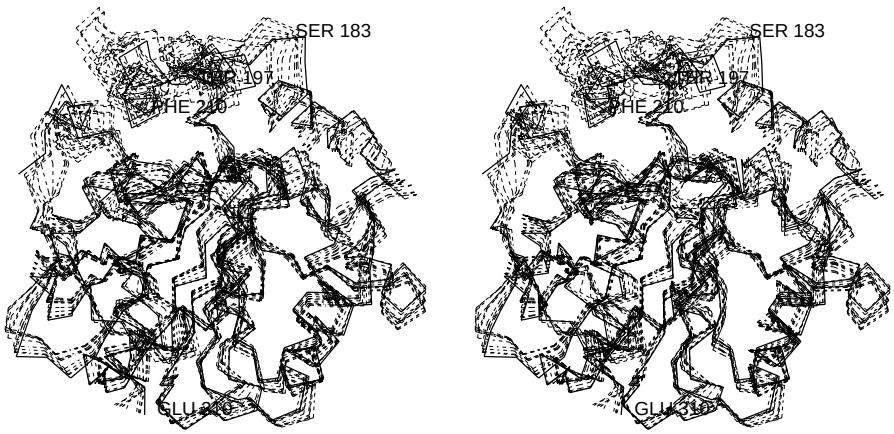

Figure 5.7: Stereoview of 11 structures from the trajectory projected onto the first 3 eigenvectors. Thick solid line: starting structure, thick dashed line: structure at 1.2 ns, thin dashed line: 9 structures separated by 120 ps. (drawn with WHATIF [48])

Finally, we also check for a possible tunnel that can act as an exit for the ion, or an entrance for the subtstrate. This was done by calculating the solvent accessible surface with the software package WHATIF [48]. We found two tunnels that are shown in fig. 5.13. There is a long narrow and curved tunnel leading to the opening created by the motion of the helix-loop-helix region and surrounded by residues Phe190, Pro182 and Phe290. Another shorter tunnel has an opening on the opposite site of the protein, surrounded by residues Lys221, Lys224 and Met225. Fig. 5.14 shows water molecules penetrating the molecule. It can be seen that water is entering through both tunnels.

5.4 Conclusions

It is not uncommon that proteins unfold during MD simulations due to inaccurate force-fields. But when unfolding takes place, in general

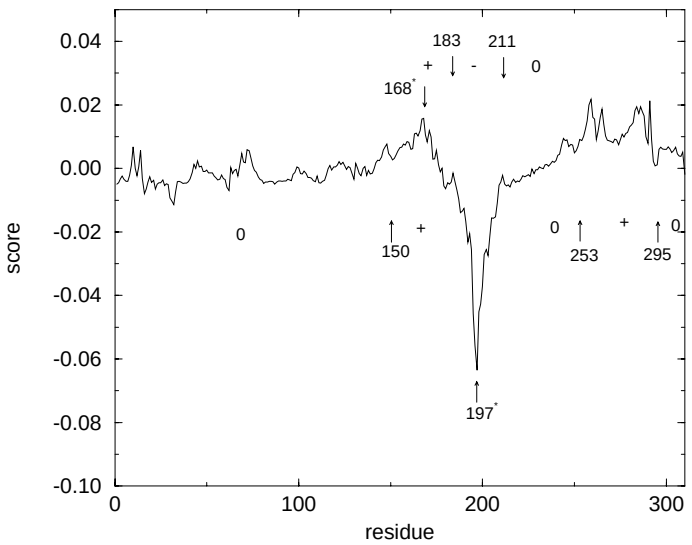

Figure 5.8: Score obtained from rigid body analysis. -, 0 and + indicate negative, neutral and positive scores referring respectively to regions A, B and C. There are two regular peaks with maxima at residues Pro168 and Thr197.

secondary structure elements start disappearing. There is no clear sign of this happening in the current simulation. Still, one has to be careful in interpreting the results above, especially because the polarization term as it was implemented, is only a rough approximation; it was originally meant to keep the chloride bound between both tryptophans. Its strong interaction with atoms that were not taken to be polarizable should be regarded with care. However, the observed large motion of the helix-loop-helix, giving rise to an opening, remains interesting. The fact that the helices mostly involved in this motion, remain not only intact, but appear to be highly stable when compared to other seconday structure elements, is certainly unusual. If these motions are assumed to be functional, the question arises how they relate to the reaction path. The chloride ion

could in principle leave the protein via the longer tunnel. This is however not very likely, as it will have to pass along several hydrophobic residues. It is more likely to take the shorter pathway, which has two positively charged residues at it's opening. The longer tunnel is more hydrophobic and leads directly to the catalytic residues. It could therefore form an entrance for the substrate.

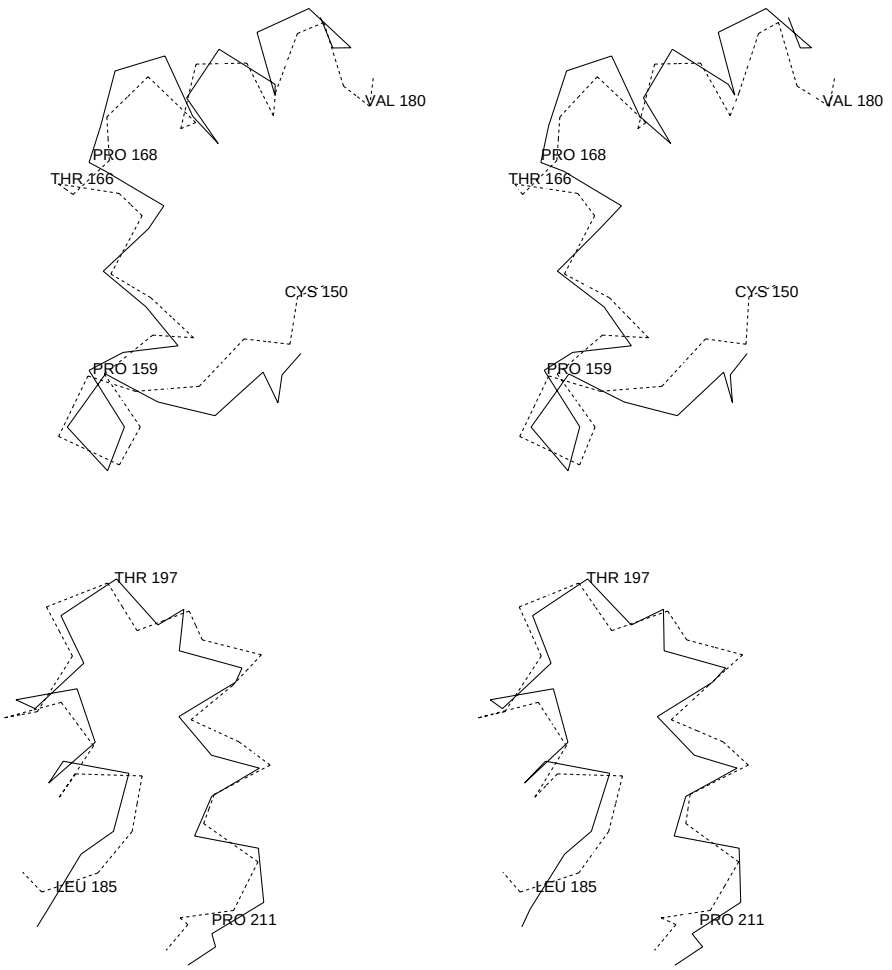

Figure 5.9: Stereoview of C_α backbone regions 150-180 (top) and 185-211 (bottom) taken from the MD structure at 1.2 ns (dotted line). Both regions are fitted to the corresponding regions from the X-ray structure (solid line) (drawn with WHATIF [48]).

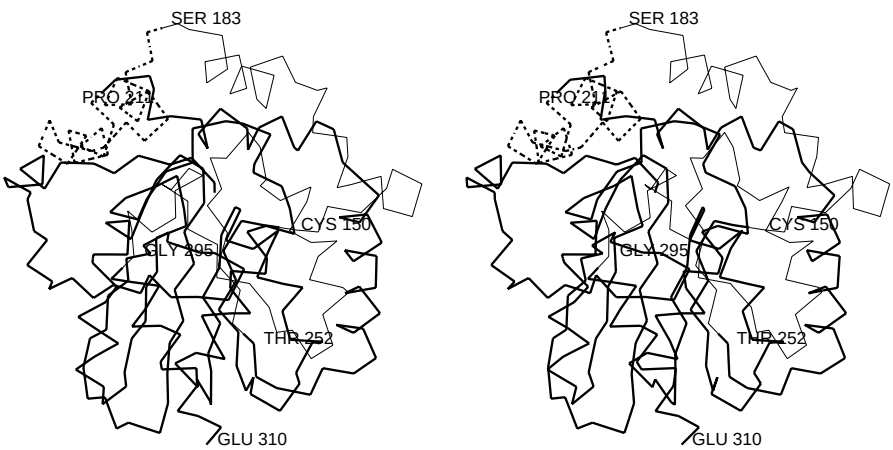

Figure 5.10: Regions A (thick dashed line), B (thick solid line) and C (thin solid line), obtained by subdividing the molecule based on rigidity scores.

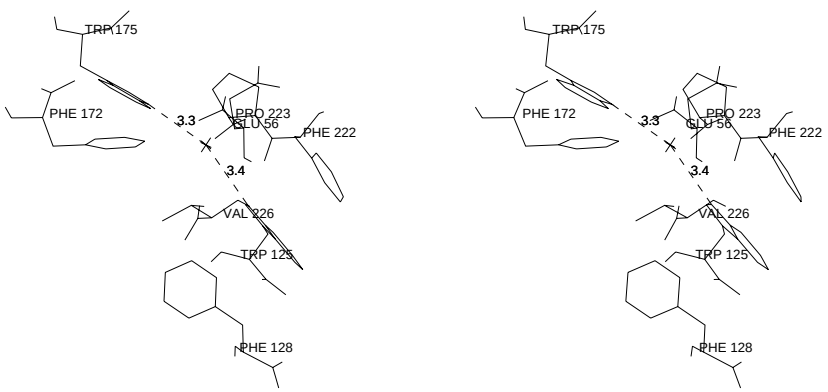

Figure 5.11: Stereoview of the chloride ion bound in the active site in the X-ray structure, the distances between the ion and $N_{\delta 1}$ of of Trp_{125} and Trp_{125} are indicated. (drawn with WHATIF [48])

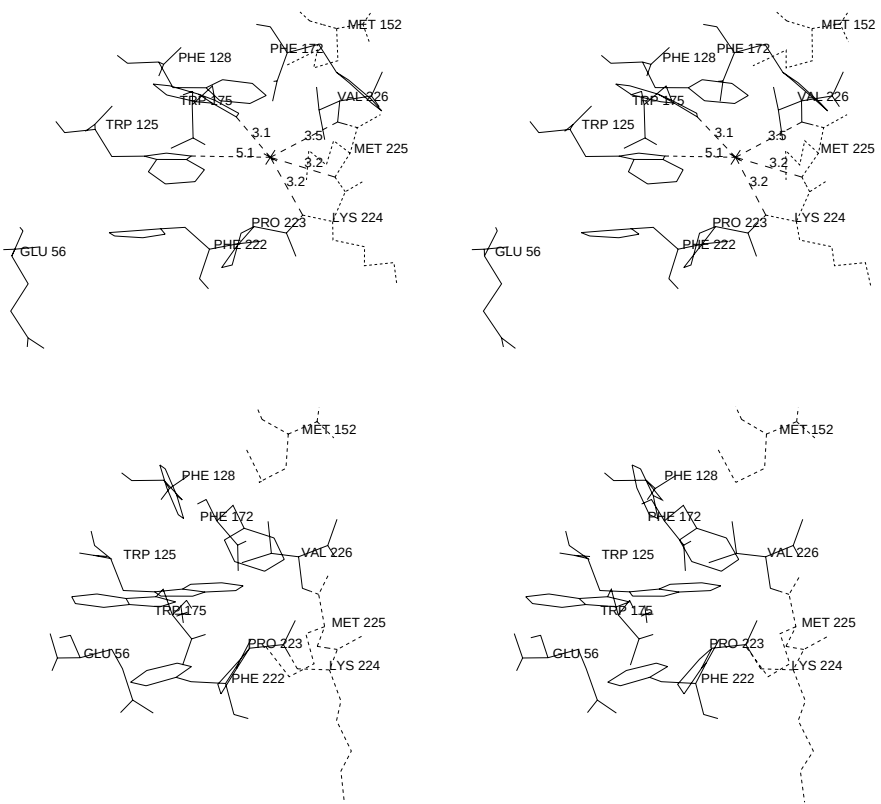

Figure 5.12: Stereoview of the environment of the chloride after 1.2 ns of simulation (top figure) and in the crystal structure (bottom figure). All polarizable residues are drawn (solid lines). In addition, all residues that,after 1.2 ns, have atoms closer to the ion then 5 Å and that were not taken to be polarizable are shown (dashed lines) (drawn with WHATIF [48]).

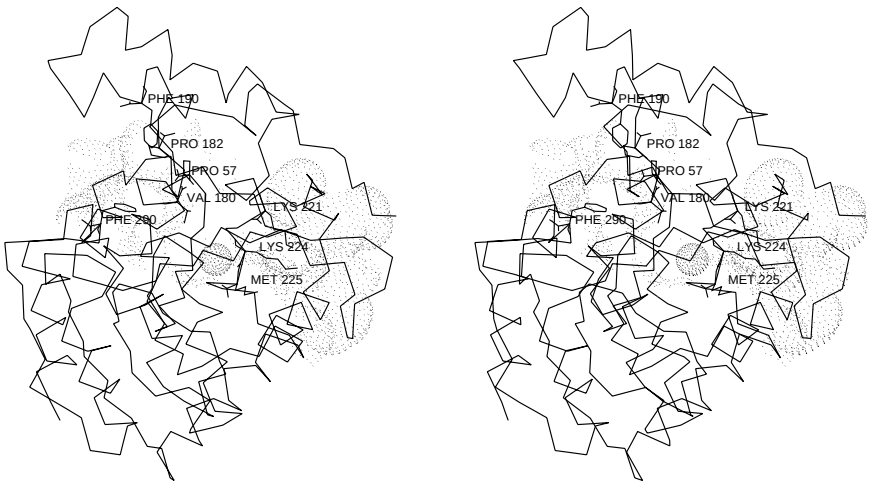

Figure 5.13: Stereoview of both tunnels leading from the active site to the surface of the protein. The chloride ion is drawn as a dotted sphere with a van der Waals radius. The longer tunnel leads to the opening on the left, surrounded by residues Phe190, Pro182 and Phe290. the shorter leads to an opening on the right-hand side, surrounded by Lys221, Lys224 and Met225 (drawn with WHATIF [48]).

Figure 5.14: Stereoview showing water penetrating the molecule. Water is entering through both tunnels shown in fig. 5.13. This picture was made by calculating the geometrical center of the protein and removing all water molecules with an oxygen distance, from this center, greater then 2.75 nm (drawn with WHATIF [48]).

Chapter 6

Suggestions for a possible cis-trans isomerisation of Pro168 in haloalkane dehalogenase

6.1 Introduction

The results obtained in chapter 5 showed an opening of the cap domain within 1 ns. If we would assume that this is the conformational transition as proposed by Schanstra et al. [14], there is a discrepancy with the experimental kinetic data. It was found in [14] that the transition $E_I \rightarrow E_{II}$ (fig. 1.4) had a rate constant of $\sim 0.3\ s^{-1}$. It is unlikely that such a process will be observed during a simulation of 1 ns. So if the MD data presented so far are reliable, there must be some other change in the protein that triggers the observed opening of the flap. It was suggested by Dr. J Damborsky [1] (personal communication) that a conformational change could involve a peptidyl-prolyl isomerisation. In fact, a similar

[1] Laboratory of Structure and Dynamics of Biomolecules, Masaryk University, Brno, Czech Republic

flap opening, accompanied by a such a cis-trans isomerisation has been observed for *Candida rugosa* lipase [49]. In this chapter we will investigate this possibility.

6.2 Which proline to consider ?

If dehalogenase can exist in two conformations and if a conformational change is indeed triggered by a proline cis-trans transition, the first question to be answered is which of the prolines is responsible. The protein contains in total 23 prolines, 7 of which are found in the cap domain. Our basic assumption was that a conformational transition would involve two (or more) regions, with strong internal interactions, but with only weak interregional interactions. For if the latter would be strong, this would probably reduce the flexibility necessarry for the transition. An investigation of the rigid bodies, as found in the previous chapters suggested two candidates. Fig. 4.9 shows Pro168 (cis) having an isolated peak, meaning it is relatively flexible. On the other hand, fig. 5.8 shows Pro211 (trans), near the edge of the opening flap. We therefore also analysed the crystal structure with the parser for protein unfolding units (PUU) [26], which was briefly introduced in chapter 2.4. Two unfolding units were thus identified, one consisting of segments 1-39, 53-65, 79-112 and 168-225, and the other containing segments 40-52, 66-78, 113-167 and 226-310. They are shown in fig 6.1 together with the cap/main domain division. It is apparent that cap and main domain do not coincide with the unfolding units as obtained by the program PUU. In fact there seems to be no relation at all between both ways of subdividing the the molecule. When the unfolding units are compared with the rigid bodies, obtained in chapter 4 (fig. 4.10) we see some correspondence.

We were now looking for a proline closest to a boundary. As can be seen from fig 6.1, Pro168 lies exactly at the boundary of both units. Some other argument that points towards this residue are the msd of the wild-type. For these reasons a possible cis-trans transition of Pro168 was

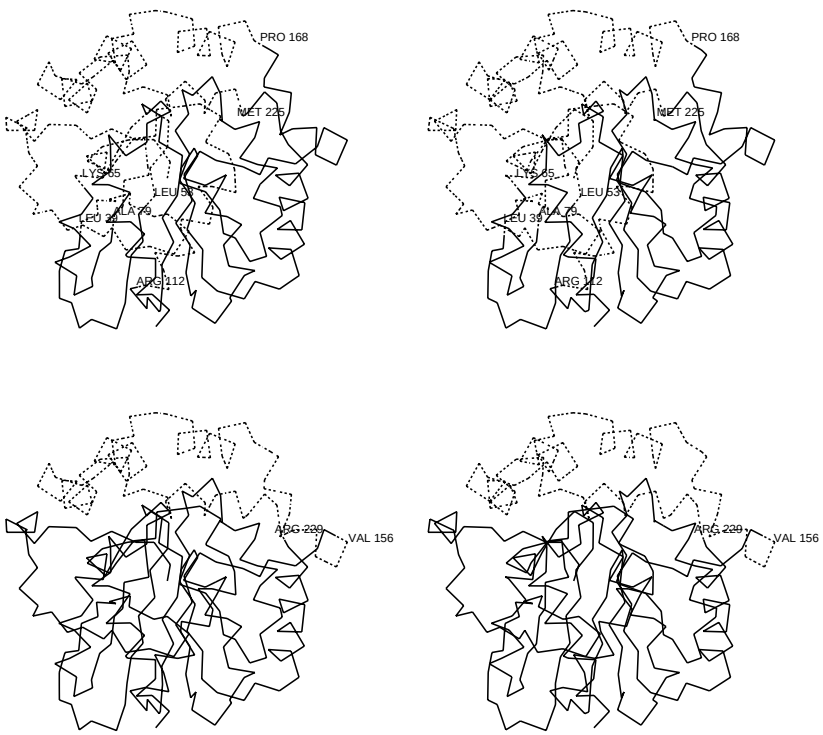

Figure 6.1: Comparison between unfolding units determined with the program *puu* (top picture) and cap and main domain (bottom figure). Residues at the boundaries are indicated. (drawn with WHATIF [48]).

first investigated.

6.3 How to show a transition?

The next question that arose was how to show that such a transition is possible. By the use of a dynamic dihedral constraint one could force an isomerisation. However, in reality it is very likely that a dihedral transition involves a globally correlated motion. I.e. a transition occurs when the protein as a whole has a proper configuration. If one forces a transition, starting from a randomly chosen structure, it could lead to an immediate unfolding. A preferable way would be to perform a simulation in which the dihedral potential of the peptide bond is removed. In this way the rate of transition will be enhanced with minimum disturbance of the system as a whole. We chose for this option.

6.4 Results

We performed two 1 ns simulations of the same systems used in both previous chapters using the same conditions. As starting configurations, the structures from those simulations after 10 ps were taken. The peptide dihedral between residues Gln167 and Pro168 was removed by defining in the topology a new dihedral with zero torsional force constant. In the following text the simulations described in the two previous chapters will be denoted as *cisdeh* and *cisdehCl* referring respectively to the simulations without and with choride. The corresponding simulation described in this chapter will be denoted as *freedeh* and *freedehCl*. Fig. 6.2 shows the behaviour of the dihedral during both simulations. In *freedeh*, large fluctuations are observed initially with a mean value of $\sim 75^\circ$ (The exact cis conformation has a dihedral angle of 0°). After $\sim$ 600 ps a transition is seen. Fluctuations become smaller and the mean value drops to $\sim 110^\circ$. In *freedehCl*, the same behaviour is observed. Now the transition

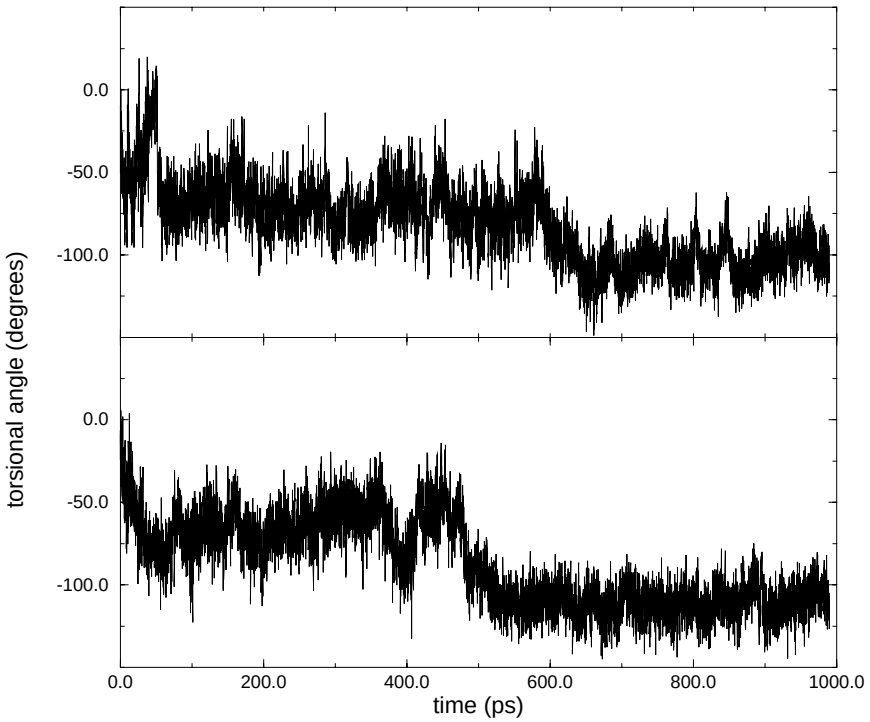

Figure 6.2: Gln167-Pro168 peptide dihedral as a function of time. Top figure: *freedeh*, Bottom figure: *freedehCl*. Exact cis and trans configurations have dihedral angles of resp. 0° and 180°

takes place at $\sim$ 500 ps. In fig 6.3 the rmsd from the crystal structure of both runs is shown. In the absence of chloride, there is an initial increase to $\sim$ 0.3 nm, whereafter a stabilization occurs. After $\sim$ 800ps there is a further increase. The value of 0.3 nm is comparable to the value obtained in chapter 4 (fig 4.3). An unusual phenomenon occurs when chloride is included: after increasing to something less than 0.3 nm the rmsd decreases to about 0.25 nm.

Figs. 6.4 and 6.5 show the behaviour of secondary structure elements. In *freedeh*, helix 4 (res. 159-166) appears to be unstable in contrast to *cisdeh*. On the other hand the short β-sheet formed by strands

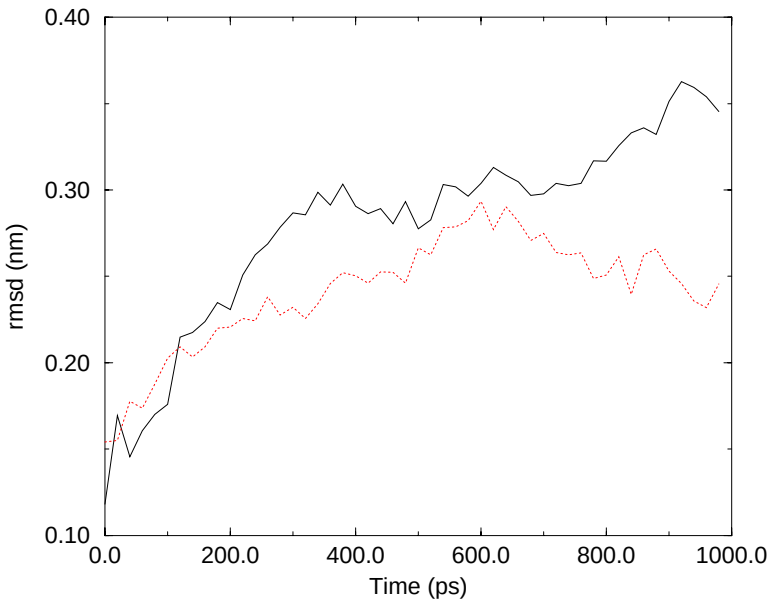

Figure 6.3: Root mean square deviation from the crystal structure as function of time. solid line: *freedeh*. dotted line: *freedehCl*

3-5 and 90-92 seems to be more stable. When comparing *freedehCl* with *cisdehCl* it can clearly be seen that the secondary structure is in general more stable. It is also interesting to note that in *trnsdehCl* the geometry of the active site remains virtually intact. In fig. 6.6 it can be seen that the chloride is still bound between Trp125 and Trp175. Also the relative positions of other residues are largely preserved.

However, were are dealing with configurations that have unphysical dihedral angles, as only angles of approximately 0° and 180° are energetically possible. To arrive at such angles, the dihedral potential had to be switched on again. This was done by slowly increasing the torsional force constant, K_{tors}, from 0 at time t_0, to it's original value K_{orig} within

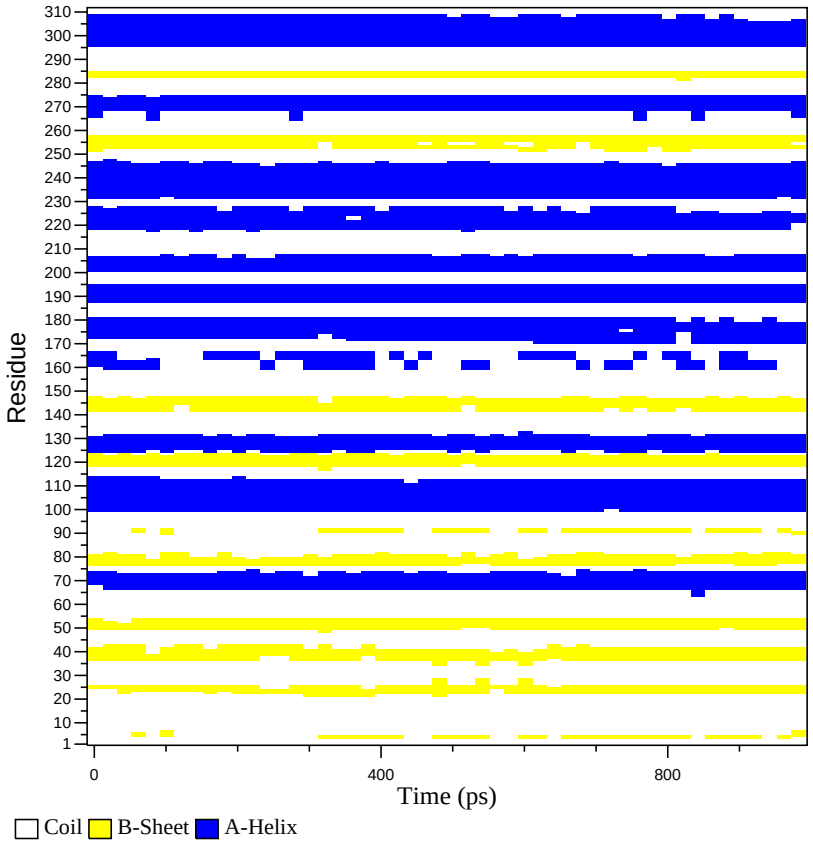

Figure 6.4: Secondary structure elements vs. time of *freedeh*

a time interval Δt according to:

$$K_{tors}(t) = K_{orig} \sin\left(0.5\pi \frac{t - t_0}{\Delta t}\right) \tag{6.1}$$

This procedure was started immediately after the transitions had occurred, i.e., t_0 was taken to be 600 ps in *freedeh*, and 500 ps in *freedehCl*. For Δt, 200 ps was chosen. After this the simulations were continued (with full torsional force constants) for another 200 ps. This resulted in total simulations of 1000 ps and 900 ps respectively (including those

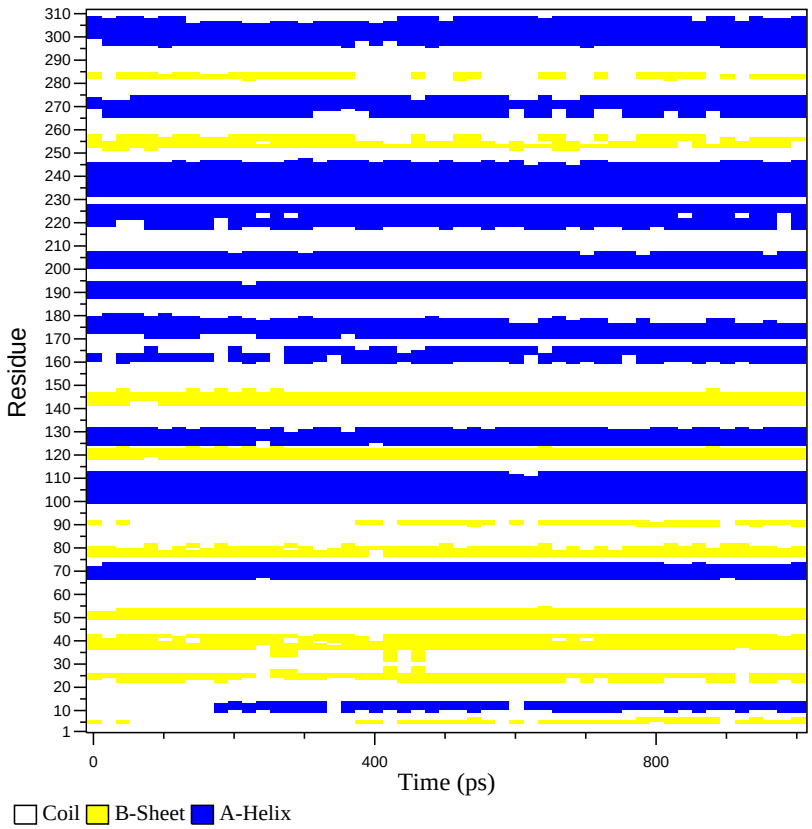

Figure 6.5: Secondary structure elements vs. time *freedehCl*

parts of the simulation with free dihedral angles before and during the transition). From now on these simulations will be denoted as *trnsdeh* and *trnsdehCl*.

In both cases the simulations resulted in trans conformations. Fig. 6.7 shows the rmsd of both simulations. It can be seen that after t_0 there is a steady increase in both simulations. The behaviour of the secondary structure elements is shown in figs. 6.8 and 6.9. In *trnsdeh* helix 4 (res. 159-166) seems to recover after t_0 but remains unstable. Also helix 5 (res.

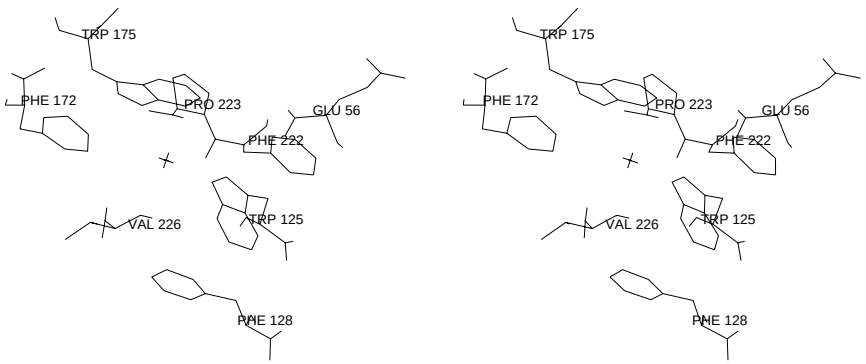

Figure 6.6: Stereoview of the active site obtained from *freedehCl* after 1 ns.

171-181, containing active site residue Trp175) becomes unstable. The overall unstability is more dramatic in *trnsdehCl*. Moreover, like in *cisdehCl*, the geometry of the active site was distorted and the chloride lost its interaction with Trp125. The B-factors, derived from both simulations are shown in fig. 6.10 together with the crystallographic B-factors. As is to be expected from all previous results, they are considerably larger then those found for *cisdeh* and *cisdehCl*. However, in *trnsdeh* the largest values are concentrated in the cap-domain, with a maximum at Thr197. This residue is situated in the opening flap that was observed in *cisdehCl*. For this reason we also performed a rigid body analysis on this simulation. The results are shown in fig. 6.11. The profile in this graph shows a striking similarity to the one in fig. 5.8, indicating that also now there is an opening of the flap. ED analysis was performed on both trajectories and fig. 6.12 shows their projections on the first 3 eigenvectors. In the top figure we can indeed recognize the motion around Thr197 as the same type of loop opening already observed in *cisdehCl*. The main-chain carbonyl group of Gln167 is directly involved in the cis-trans transition. In none of the final structures obtained from the simulations it is involved in an intramolecular hydrogen bond. Neither in the crystal structures is this interaction observed.

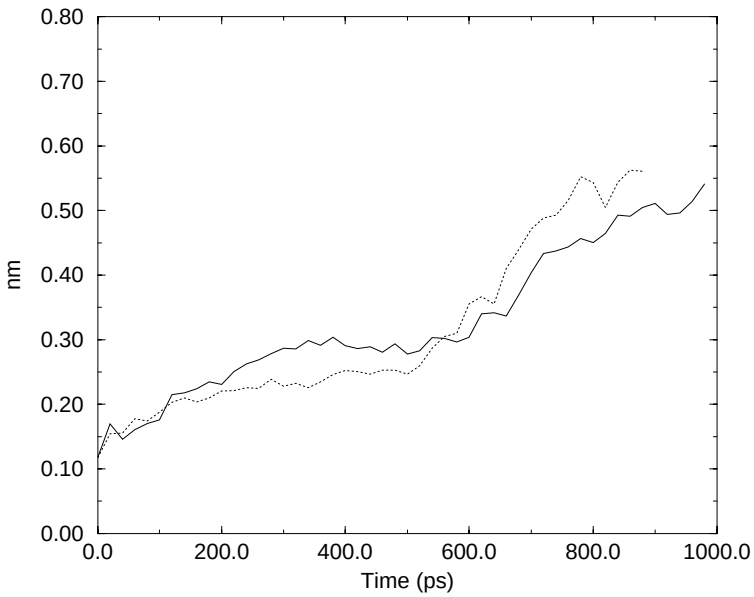

Figure 6.7: rmsd from crystal structure, for *trnsdeh* (solid line) and *trnsdehCl* (dotted line). t_0 was 600 ps and 500 ps respectively (see text).

Finally, a simulation of the system containing no chloride in the active site was again performed, but now the dihedral potential of the peptide bond of Pro211 was removed. This system unfolded rapidly as can be seen from the secondary structure elements in fig. 6.13. This calculation was stopped after 750 ps. Similar results were obtained when the Pro223 peptide dihedral potential was removed.

6.5 Conclusions

It is remarkable to find that when a torsional potential of one of its peptide bonds is removed, a protein does not unfold. It is even more so

if the configuration appears to become more stable, as was observed in *freedehCl*. The fact that the geometry of the active site doesn't change dramatically gives confidence in the model that was introduced to include polarization effects. When the dihedral potential is slowly reintroduced, both systems showed considerable structural changes. This may be due to inaccuracies of the force field, but also the method of reintroducing this potential may be responsible. As was already mentioned, cis-trans transitions not just involve one peptide bond, but are probably the result of the overall configuration of the protein at a certain time. In the physical world, cis-trans transitions do not occur frequently because of the high energy barrier that has to be overcome. But when such a transition takes place, it is a fast process during which the change of the overall configuration will be small. In our simulations the process was very slow, and came with significant configurational changes. An interesting feature concerns the flap opening. While the flap opens in *trnsdeh* , no such behaviour is observed in *trnsdehCl*. This is in contrast with the results obtained from *cisdeh* and *cisdehCl* where the opening was found only in the presence of chloride. So it not unrealistic to assume that chloride has significant effects on the dynamics of dehalogenase. Some possible implications of this will be discussed in the next chapter.

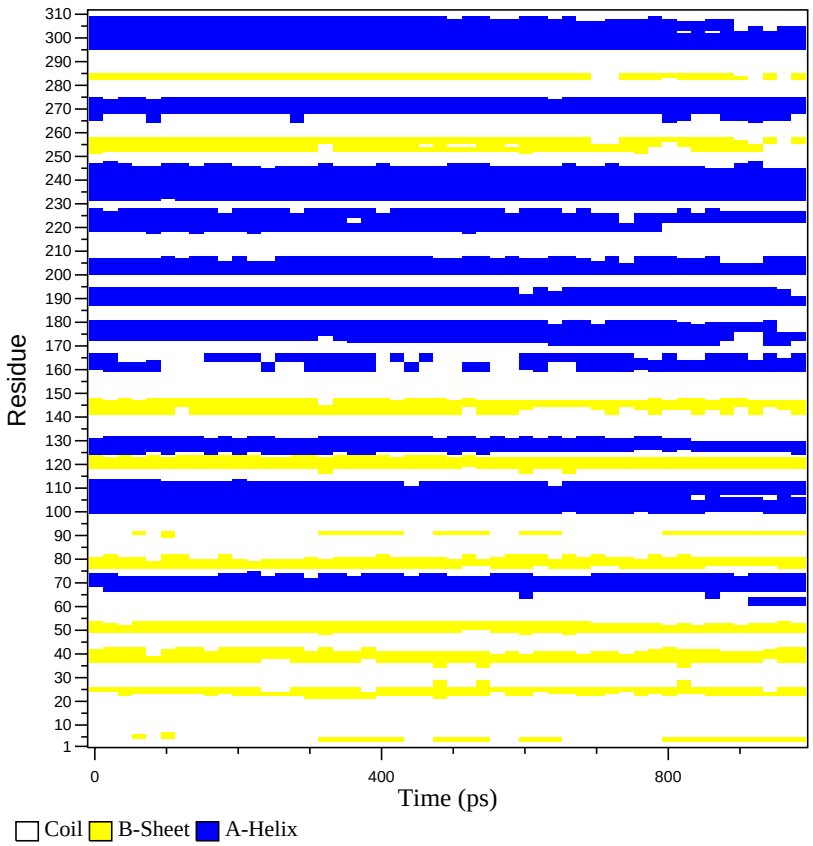

Figure 6.8: Secondary structure elements vs. time of *trnsdeh*. The dihedral potential was slowly reintroduced in the time interval between 600-800ps (see text).

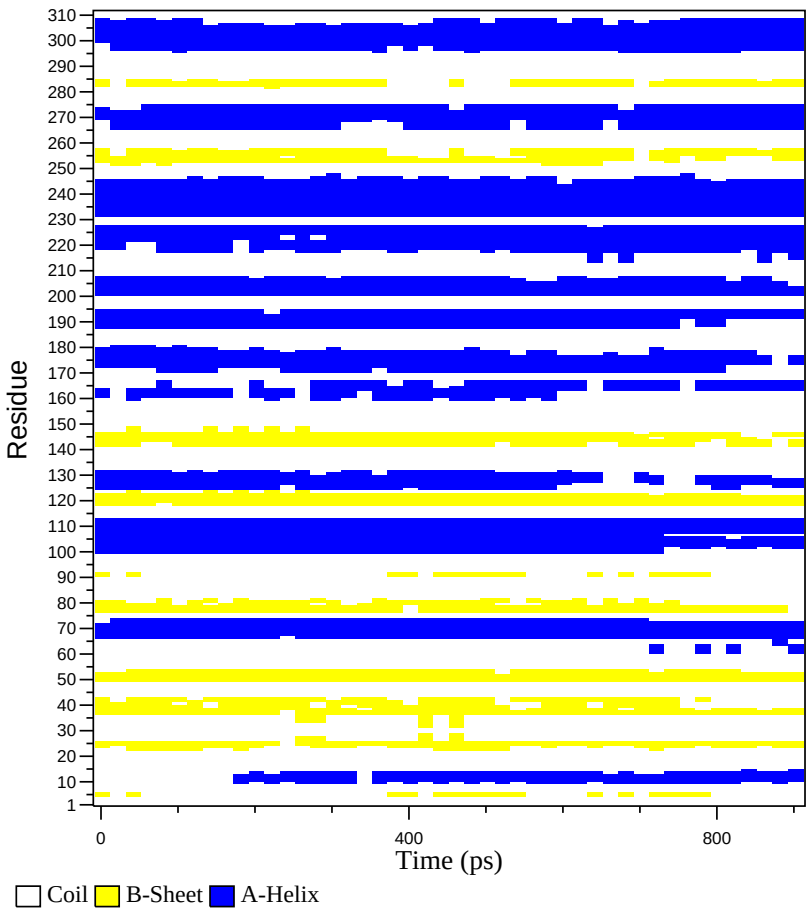

Figure 6.9: Secondary structure elements vs. time of *trnsdehCl*. The dihedral potential was slowly reintroduced in the time interval between 500-700ps (see text).

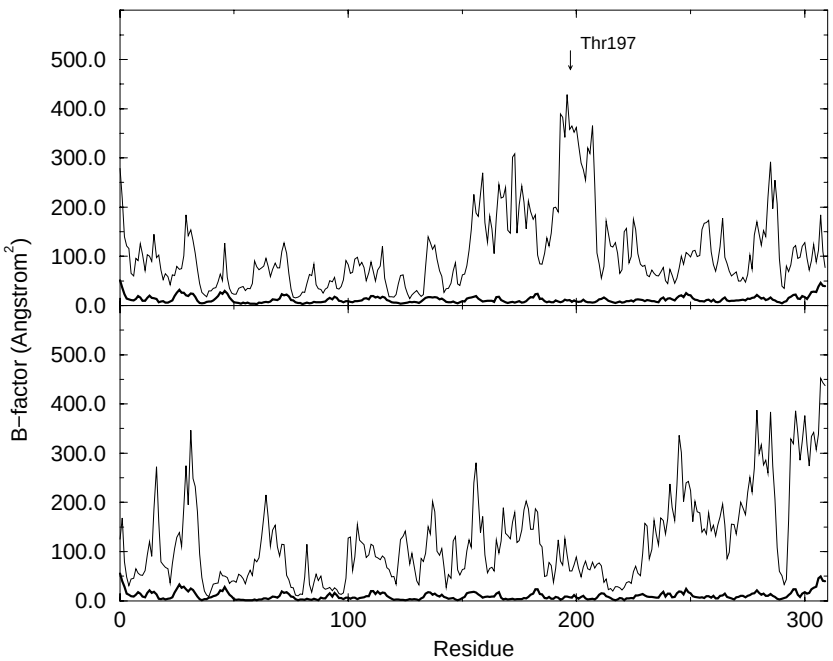

Figure 6.10: B-factors derived from the simulations (thin line) together with the crystallographic B-factors (thick line). Top figure: *trnsdeh* bottom figure *trnsdehCl*.

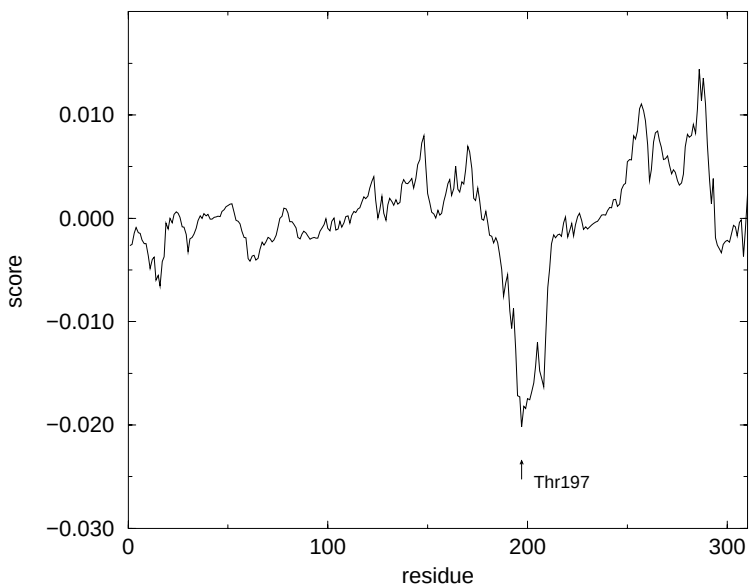

Figure 6.11: Scores obtained from rigid body analysis of *trnsdeh*.

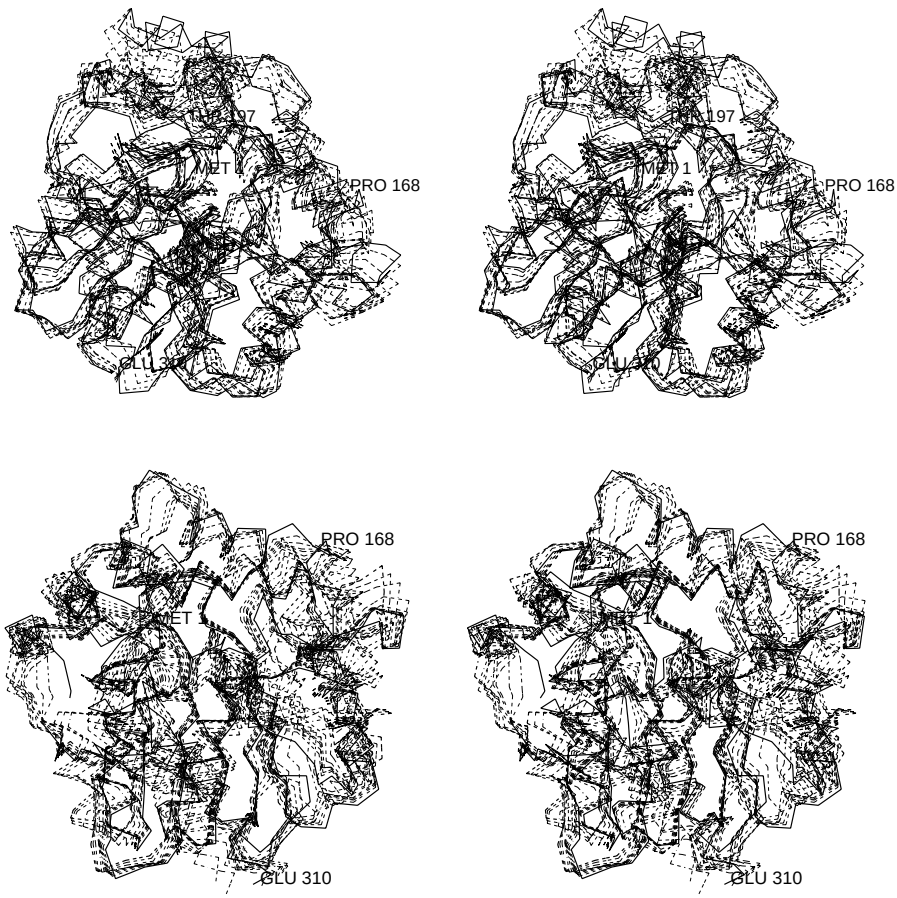

Figure 6.12: Trajectories projected on their first three eigenvectors obtained from ED analysis. Thick solid lines show starting structures, thick dashed lines final structures, thin dashed lines show 9 structures separated by equal time intervals. Top figure: *trnsdeh*, structures are separated by 100 ps. Bottom figure: *trnsdehCl* structures are separated by 90 ps.

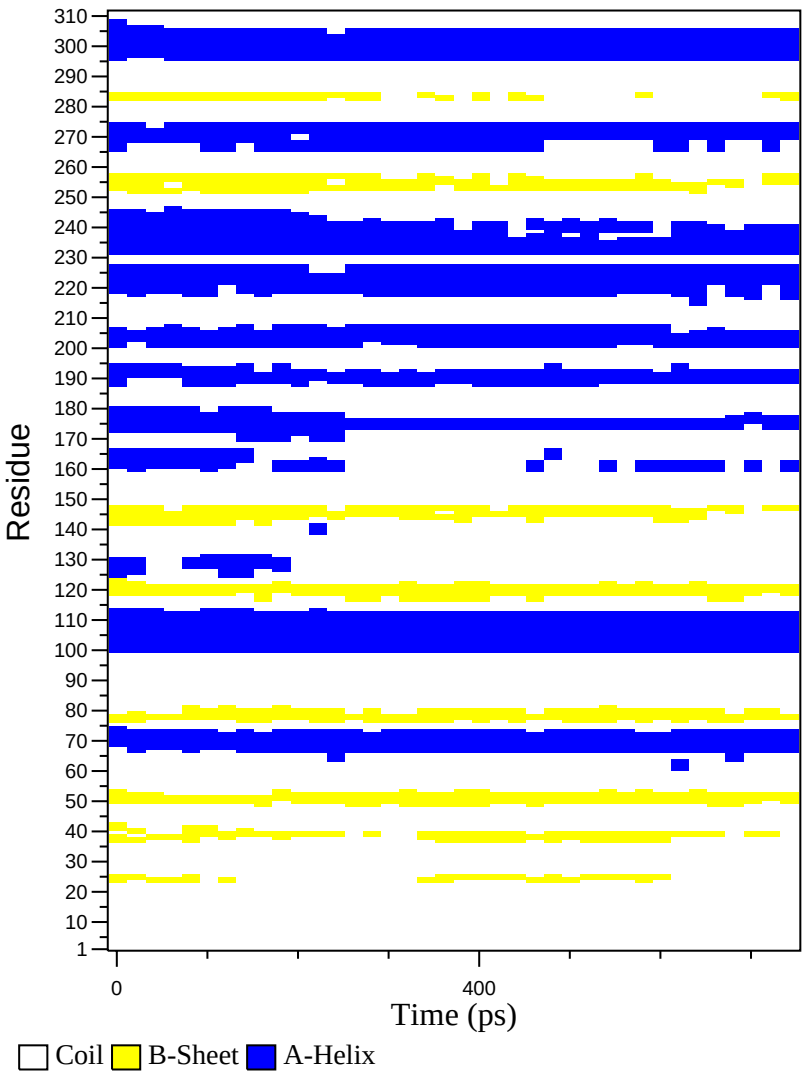

Figure 6.13: Secondary structure elements vs. time of the simulation without chloride, with the dihedral potential of the peptide bond of Pro211 removed.

Chapter 7

General conclusion

The results described in the previous chapters may shed new light on the kinetics of halide release of dehalogenase. In this chapter some implications of a possible Pro168 cis-trans isomerization will be discussed.

In the following text we will call dehalogenase in the cis-conformation E_{cis}, and in the trans conformation E_{trans} and their corresponding halide-bound states $E_{cis}.X$ and $E_{trans}.X$. We can summarize the results obtained from the MD simulations so far as follows:

- E_{cis} has large structural fluctations, but appears to be stable with respect to its X-ray structure
- $E_{cis}.X$ shows a large motion in the amino acid sequence region 185-211, giving rise to an opening of the active site and resulting in two tunnels that connect the active site with the exterior of the protein.
- Dihedral transitions $E_{cis} \longrightarrow E_{trans}$ and $E_{cis}.X \longrightarrow E_{trans}.X$ are observed when the dihedral potential of the peptide bond between Gln167 and Pro168 is removed.
- Although E_{trans} shows large structural changes during the simulation, one can identify the same cap-domain motions as found for $E_{cis}.X$

- Also after the transition $E_{cis}.X \longrightarrow E_{trans}.X$ substantial structural changes are observed, but now there appears to be no tendency towards an opening of the cap domain $E_{trans}.X$

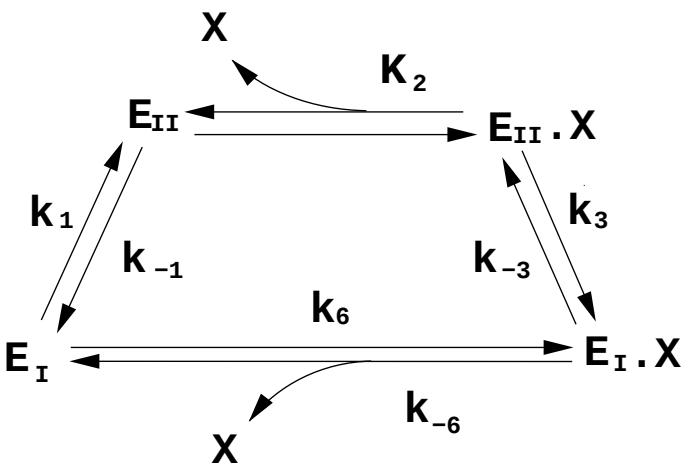

Figure 7.1: Kinetic scheme suggested for chloride release [14]. $E_I.X$ is chloride complex formed after catalysis. The upper route is predominant at low chloride concentration. The upper route at high halide concentration. Kinetic constants are given in table 7.1.

Kinetic studies [14] have revealed that halide release can occur via two routes. Fig. 7.1 shows the scheme as it was suggested for chloride release, where $E_I.X$ is the enzyme-halide complex produced by the catalysis reaction. Kinetic constants are given in table 7.1. The lower route is believed to involve rapid formation of a collision complex followed by a slow enzyme isomerization. It prevails at high halide concentrations. In the upper route, fast halide binding is preceeded by a slow conformational change from E_I to E_{II}. The latter is predominant at lower halide

concentration.

k_1	$3\pm0.3\ s^{-1}$
k_{-1}	$>300\ s^{-1}$
k_3	$>1450\ s^{-1}$
k_{-3}	$14.5\pm0.5\ s^{-}1$
k_6	$0.0085\pm0.005\ mM^{-1}.s^{-1}$
k_{-6}	$0.66\pm0.03\ s^{-1}$
K_2	-

Table 7.1: Kinetic data for fig. 7.1 [14].

If we try to combine the MD results with the kinetic data and assume that a cis-trans isomerization is involved in halide release, we arrive at three possible ways of doing so.

possibility 1: Because in the crystal (cis) conformation all reaction steps, except for halide release, can take place [9], the obvious conclusion would be that E_I corresponds to E_{cis} and E_{II} to E_{trans}. In combination with the MD results this assumption would convert fig. 7.1 into fig. 7.2. This scheme implies that the structure opens after catalysis has taken place. The faster (upper) halide release route, would now involve a cis to trans isomerization combined with a closure of the flap which is followed by a fast dissociation. This means that halide release from an open structure, the lower route, is slower than that from a closed structure. Although this might seem implausible, our MD data do not exclude this option. In the open structure the halide ion might obtain more favourable interactions with the enzyme, that were not observed during the simulation because of its small timescale and/or the inaccuracy of the forcefield.

possibility 2: If we assume that E_I and E_{II} correspond to E_{trans} and E_{cis} we obtain the scheme shown in fig. 7.3. Now new questions arise. From k_1 and k_{-1} in table 7.1 we learn that E_I and E_{II} exist in

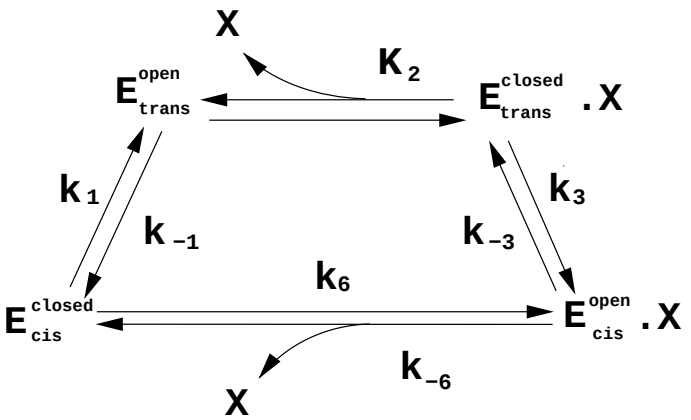

Figure 7.2: Adjusted scheme were E_I and E_{II} are assumed to correspond to E_{cis} and E_{trans} respectively. $E_{cis}.X$ is the product from catalysis

equilibrium:

$$E_I \overset{K_{ct}}{\rightleftharpoons} E_{II} \tag{7.1}$$

with $K_{ct} \ll 0.01$. In the present case this would mean that the trans conformation is in fact the most abundant conformation in solution. Up till now, no X-ray structure of this conformation has been found.

According to the MD results, $E_{cis}.X$ will, unlike its X-ray structure, adopt an open conformation. However, crystallization experiments were all performed in the absence of halide. The halide bound structure was obtained by soaking crystals in halide solution [9]. In this way, a flap-opening is inhibited. During the simulation E_{cis} doesn't show large conformational changes with respect to the X-ray structure, as is to be expected. Furthermore, under equal conditions, it is very likely that one isomer crystallizes more readily than the other. E_{cis} has the most compact conformation, that may allow for more favorable crystal packing arrangements. It is also very probable that the open structure will be less rigid than the closed one, meaning that in solution the latter will have less entropy. So crystallization of E_{cis} will be favored because it is ac-

companied by a smaller entropy decrease relative to E_{trans}. So, although E_{trans} may be the abundant conformation, E_{cis} might actually crystallize. Scheme 7.3 is also supported by the fact that, when the dihedral

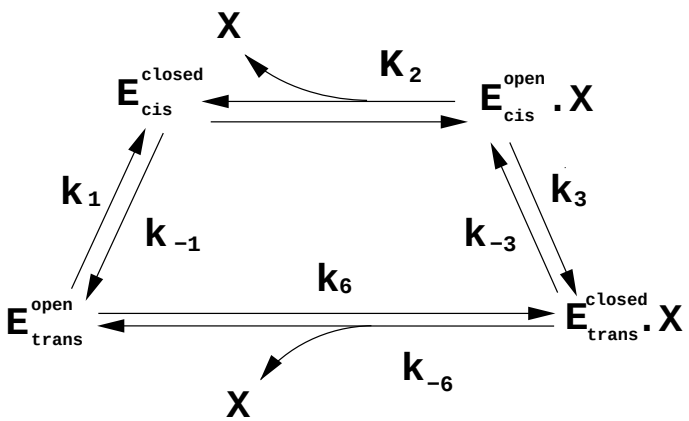

Figure 7.3: adjusted kinetic scheme based on MD results and the assumption that E_{trans} and E_{cis} correspond respectively to E_I and E_{II}. $E_{trans}.X$ is the product from catalysis.

potential of the peptide bond between Gln167 and Pro168 is removed, E_{cis} as well as $E_{cis}.X$ tend to adopt a trans-conformation.
Another diffulty arises from the observation that in the crystal (cis) conformation catalysis still occurs [9]. As a consequence, our new scheme implies that both conformations must be catalytically active.

possibility 3: Finally it is also possible that the cis-trans isomerization observed during the simulations corresponds to the lower route. In this case two isomerizations have to occur: cis $\longrightarrow$ trans before halide release, and trans $\longrightarrow$ cis afterwards. Experimentally only one isomerization could be detected.

Some remarks must be made on the first two possibities. Recently,

kinetic studies have been performed on the mutant Pro168 $\longrightarrow$ Ser (G.H. Krooshof & D.B. Janssen, not yet published, personal communication). This mutant showed an increase in the ratio k_1/k_{-1} while a decrease in the ratio k_3/k_{-3} was found. This means that conformation E_I has become thermodynamically less favorable relative to E_{II}. It is known that prolines in general destabilize the trans relative to the cis conformation in peptide bonds [50]. This would make possibility 1 more probable. But it remains difficult to prefer one possibility over the other. Further kinetic studies of mutants may provide the final answer.

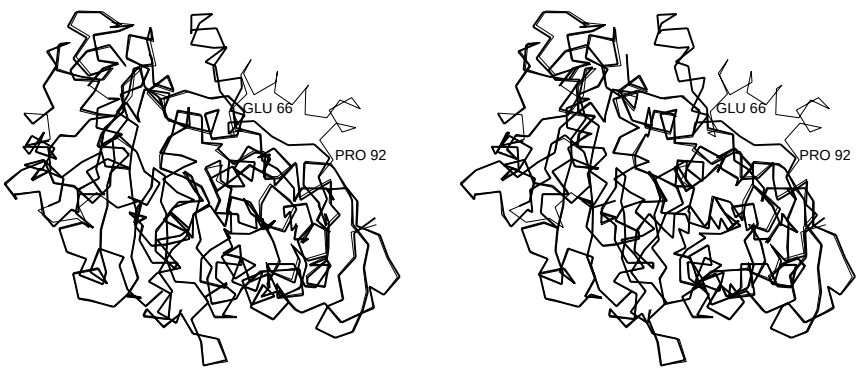

Figure 7.4: Two stuctures of *Candida rugosa* lipase, 1CRL (open) and 1THR (closed), superimposed. Like in [49] the tranformation matrix was calculated based on all C_α atoms except residues 61-96. Thin line: 1CRL, thick line: 1THR. (drawn with WHATIF [48])

A flap opening, triggered by a proline cis-trans isomerization, has been observed for *Candida rugosa* lipase [49], like dehalogenase a α/β-hydrolase fold protein. Two X-ray structures were found, designated as open and closed conformations. Fig. 7.4 shows both structures (Brookhaven Data Bank [36], entries 1CRL and 1THR, respectively open and closed structures). There are clearly defined hinge points at Glu66 and Pro92. In the open and closed structures, the peptide bond between Ser91 and

Pro92 has respectively a cis- and trans-conformation. It is however unlikely that both mechanisms have a common ancestor. The MD results for dehalogenase suggest that opening and closing of the flap is not only determined by Pro168 peptide conformation, but also by the presence or absence of a halide ion in the active site. In addition, Pro168 forms no part of the flap itself, the effect of a cis-trans transition on the flap is transmitted through the backbone over more than 15 residues as the flap has its closest hinge point at approximately Leu185. In *Candida rugosa* lipase the peptide bond between Ser91 and Pro92 forms one of the hinge points. So, there is clearly a difference in mechanism.

At present, MD still has its shortcomings. The timescale of protein simulations, presently restricted to several nanoseconds, is small compared to the timescale in which proteins act. In addition, the forcefields have only a limited accuracy. Much progress is still being made in order to improve these aspects of the technique. But even at the present state of the art, by the use MD, combined with the proper analysis tools, one can study properties that are not accessible to any experimental technique and thereby produce new points of view. The author hopes that this thesis offers a good example.

Chapter 8

Samenvatting

Enzymen vormen een van de belangrijkste klassen van verbindingen die voorkomen in de natuur en spelen een rol als katalysator van tal van biochemische processen. Haloalkaan Dehalogenase is zo'n enzym. Het komt voor in een bacterie-stam, genaamd *Xanthobacter autotrophicus* GJ10, die in staat is 1,2-dichloorethaan te gebruiken als zowel energie- als koolstofbron. Haloalkaan dehalogenase katalyseert een van de stappen in het afbraakproces: de substitutie van één van de chlooratomen door een hydroxylgroep. Röntgendiffractie-experimenten hebben een gedetailleerd beeld gegeven van het reactiemechanisme en kinetiekstudies hebben laten zien dat de laatste stap in dit proces, de dissociatie van het enzym-chloridecomplex, snelheidsbepalend is.

Dit proefschrift beschrijft de resultaten van verscheidene Moleculaire Dynamica simulaties (MD) van haloalkaan dehalogenase, hierna kortweg dehalogenase genoemd, waarbij met name onderzocht is wat de invloed is van de aanwezigheid van een chloride-ion in het actieve centrum. Er wordt ook ingegaan op enkele recent ontwikkelde MD analysetechnieken.

In *hoofdstuk 1* wordt een beeld gegeven van experimenteel werk dat tot nu toe is uitgevoerd aan dehalogenase. Er wordt ingegaan op de kristalstructuur en het reactiemechanisme. Daarnaast wordt aandacht

besteed aan enkele kinetiekexperimenten met betrekking tot de dissociatie van het enzym-chloridecomplex. Deze experimenten suggereren dat dissociatie samengaat met een conformatieverandering van het eiwit.

Hoofdstuk 2 concentreert zich op enkele rekentechnische aspecten. Naast een summiere uitleg van de technieken die gebruikt worden in standaard simulaties, wordt een toevoeging aan het MD krachtenveld beschreven: de *polariseerbaarheid*. In het kort gezegd is een krachtenveld de verzameling formules en parameters die bepalen wat de interacties zijn tussen atomen. Een van de termen die in het gebruikte krachtenveld miste, was de zogenaamde polarisatieterm: wanneer een atoom in een elektrisch veld geplaatst wordt zal zijn elektronenwolk vervormen waardoor een dipool wordt geïnduceerd. Een gevolg hiervan is dat ook elektrisch ongeladen atomen een elektrostatische interactie met hun omgeving kunnen ondervinden. Deze interactie speelt vooral een rol wanneer ionen in een apolaire omgeving worden geplaatst, zoals een chloride-ion in het actieve centrum van dehalogenase. Een model om deze interactie, met een zekere nauwkeurigheid, toe te voegen aan het krachtenveld wordt uitgewerkt. Ook wordt ingegaan op de *Essentiele Dynamica* (ED) analysetechniek. Hoewel dit het onderwerp is van het volgende hoofdstuk wordt aan de hand van twee eenvoudige systemen een eerste indruk gegeven waarbij de wiskunde voor een breder publiek toegankelijk is gehouden. Als laatste wordt de *Rigid Body* analyse toegelicht. Het gaat hierbij om, met behulp van MD trajectoria, een beeld te krijgen van de aanwezigheid van starre regio's in het eiwit.

In *hoofdstuk 3* wordt de ED techniek verder uitgediept en toegespitst op eiwitten. De structuur van een eiwit, bestaande uit N atomen, kan worden beschreven in een driedimensionale ruimte, waarbij de positie van ieder atoom is bepaald door drie coordinaten. We kunnen zo'n molecuul echter ook beschrijven in een $3N$-dimensionale ruimte. In deze ruimte wordt de volledige structuur beschreven door één punt. Een MD simulatie resulteert in een verzameling structuren als functie van de tijd. Deze worden nu gerepresenteerd als een verzameling punten. ED analyse is in

principe niets anders dan Principale Componenten Analyse. Dit betekent dat er in de $3N$-dimensionale ruimte een nieuw assenstelsel wordt geconstrueerd, zodanig dat er met zo weinig mogelijk assen zoveel mogelijk beweging van het eiwit wordt beschreven. Wanneer deze methode wordt toegepast op een MD trajectorium van lysozym, blijkt dat er bewegingen kunnen worden herkend die geconcentreerd zijn rondom het actieve centrum. Deze bewegingen lijken daardoor essentieel voor de functie van het eiwit.

Hoofdstuk 4 en 5 behandelen achtereenvolgens MD simulaties van dehalogenase, zonder en met een chloride-ion in het actieve centrum. In beide gevallen valt op dat het eiwit een grote flexibiliteit vertoont. Met name het cap-domein, een van beide domeinen van het eiwit waartussen zich het actieve centrum bevindt, laat veel beweging zien. Maar het meest opvallend zijn twee naast elkaar gelegen helices die, in de simulatie met chloride, als een flap openen.

Tijdens het onderzoek werd door een van de medewerkers gesuggereerd dat een conformatieverandering veroorzaakt kon worden door een cis-trans isomerisatie van een proline tweevlakshoek. Dit is het onderwerp van *hoofdstuk 6*. Uitgaande van een dergelijk proces wordt onderzocht om welke proline het zou kunnen gaan. Ook worden berekeningen beschreven waarbij een cis-trans isomerisatie van Pro168 wordt gesimuleerd. Dit wordt gedaan aan dehalogenase zowel in afwezigheid als in aanwezigheid van chloride. Nu blijkt dezelfde flapopening als eerder beschreven plaats te vinden, maar ditmaal alleen in afwezigheid van chloride.

In *hoofdstuk 7* wordt een poging gedaan om de resultaten van de simulaties in overeenstemming te brengen met de experimentele resultaten. Er zijn in principe drie mogelijkheden die elk worden beschreven. Welke van deze mogelijkheden het meest waarschijnlijk is, valt momenteel moeilijk te voorspellen. Kinetische experimenten met mutanten zullen hierover in de toekomst uitsluitsel moeten geven.

MD kent zijn beperkingen. Zo is de tijdschaal waarop gesimuleerd wordt, voor eiwitten beperkt tot enkele nanoseconden, kort in vergelijking met de tijdschaal waarbinnen eiwitten een volledige reactie cyclus uitvoeren. Daarnaast bezitten de gebruikte krachtenvelden een beperkte nauwkeurigheid. Er wordt nog steeds veel werk verricht om beide tekortkomingen te verbeteren. Maar ook uitgaande van de huidige stand van zaken biedt MD mogelijkheden die voorheen ongekend waren. Samen met de juiste analysetechnieken kunnen eigenschappen worden bestudeerd die met de huidige experimentele technieken niet meetbaar zijn. Met dit proefschrift hoopt de auteur hiervan een goed voorbeeld te hebben gegeven.

Appendix A

The occurrence of linear near constraints as eigenvectors of the covariance matrix

In this appendix we show that every exact or approximate holonomic constraint is associated with an exactly or approximately zero eigenvalue of the covariance matrix.

If we have P holonomic constraints in our system we may express them using an implicit form for the relation between Cartesian coordinates:

$$\mid G_i(\boldsymbol{x}) \mid \leq \epsilon \qquad i = 1, 2, ...P \tag{A.1}$$

with $\epsilon \geq 0$. If these constraints can be linearized around the average position of $\boldsymbol{x}$, we can replace the previous inequality by

$$\mid G_i(\langle \boldsymbol{x} \rangle) + \nabla G_i \cdot (\boldsymbol{x} - \langle \boldsymbol{x} \rangle) \mid \leq \epsilon \tag{A.2}$$

where ∇G_i is taken at $\langle \boldsymbol{x} \rangle$. Since

$$\mid G_i(\langle \boldsymbol{x} \rangle) \mid \leq \epsilon \tag{A.3}$$

we must have

$$\mid \nabla G_i \cdot (\boldsymbol{x} - \langle \boldsymbol{x} \rangle) \mid \leq 2\epsilon \tag{A.4}$$

If we multiply the covariance matrix at the right side by ∇G_i we

$$\begin{aligned} C\nabla G_i &= \langle(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)^T\rangle\nabla G_i \\ &= \langle(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)^T\nabla G_i\rangle \\ &= \langle(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)[\nabla G_i \cdot (\boldsymbol{x} - \langle\boldsymbol{x}\rangle)]\rangle \end{aligned} \tag{A.5}$$

Because we have near constraints, ϵ tends to zero, hence

$$\begin{aligned} \lim_{\epsilon\to 0} C\nabla G_i &= \lim_{\epsilon\to 0}\langle(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)[\nabla G_i \cdot (\boldsymbol{x} - \langle\boldsymbol{x}\rangle)]\rangle \\ &= 0 \end{aligned} \tag{A.6}$$

From equation A.6 it follows that every linear combination of gradients ∇G_i tends to be an eigenvector with an almost zero eigenvalue of C. Thus every time we have in a system P linear near constraints as defined by equation A.2, then we will always obtain P corresponding eigenvectors with almost zero eigenvalues from C.

Appendix B

The relation between the number of configurations and the number of non-zero eigenvalues

In this appendix it will be shown that a covariance matrix, constructed from S configurations cannot have more than $S-1$ eigenvectors with nonzero eigenvalues.

The covariance matrix C is given by:

$$C = \frac{1}{S}\sum_i \Delta \boldsymbol{x}_i \Delta \boldsymbol{x}_i^T \tag{B.1}$$

with $i = 1, 2, ...S$ and $\Delta \boldsymbol{x}_i$ is the displacement vector of configuration i given by:

$$\Delta \boldsymbol{x}_i = \boldsymbol{x}_i - \frac{1}{S}\sum_j \Delta \boldsymbol{x}_j \tag{B.2}$$

where $j = 1, 2...S$ and $\boldsymbol{x}_j$ is the position vector of configuration j which contains all the position coordinates of the system. If we take the sum

of all S displacement vectors we obtain:

$$\begin{aligned} \textstyle\sum_i(\boldsymbol{x}_i - \frac{1}{S}\sum_j \boldsymbol{x}_j) &= \sum_i \boldsymbol{x}_i - \sum_i(\frac{1}{S}\sum_j \boldsymbol{x}_j) \\ &= \textstyle\sum_i \boldsymbol{x}_i - S\frac{1}{S}\sum_j \boldsymbol{x}_i \\ &= \mathbf{0} \end{aligned} \tag{B.3}$$

which means that we can expres one displacement vector as a linear combination of the others:

$$\Delta x_i = -\sum_{j\neq i} \Delta \boldsymbol{x}_j \tag{B.4}$$

where the summation contains $S-1$ terms. We can separate one of the terms in equation B.1 from the summation to obtain:

$$C = \frac{1}{S}\Delta \boldsymbol{x}_i \Delta \boldsymbol{x}_i^T + \frac{1}{S}\sum_{j\neq i} \Delta \boldsymbol{x}_j \Delta \boldsymbol{x}_j^T \tag{B.5}$$

where the summation also contains $S-1$ terms. Substituting equation B.4 into equation B.5 gives:

$$\begin{aligned} C &= \tfrac{1}{S}(-\sum_{k\neq i}\Delta \boldsymbol{x}_k)(-\sum_{l\neq i}\Delta \boldsymbol{x}_l)^T + \tfrac{1}{S}\sum_{j\neq i}\Delta \boldsymbol{x}_j \Delta \boldsymbol{x}_j^T \\ &= \tfrac{1}{S}(\sum_{k\neq i}\Delta \boldsymbol{x}_k)(\sum_{l\neq i}\Delta \boldsymbol{x}_l)^T + \tfrac{1}{S}\sum_{j\neq i}\Delta \boldsymbol{x}_j \Delta \boldsymbol{x}_j^T \\ &= \tfrac{1}{S}\sum_{k\neq i}[\Delta \boldsymbol{x}_k(\sum_{l\neq i}\Delta \boldsymbol{x}_l)^T] + \tfrac{1}{S}\sum_{j\neq i}\Delta \boldsymbol{x}_j \Delta \boldsymbol{x}_j^T \\ &= \tfrac{1}{S}\sum_{j\neq i}[\Delta \boldsymbol{x}_j(\Delta \boldsymbol{x}_j + \sum_{l\neq i}\Delta \boldsymbol{x}_l)^T] \\ &= \tfrac{1}{S}\sum_{j\neq i}\Delta \boldsymbol{x}_j(\Delta \boldsymbol{x}_j + \sum_{l\neq i}\Delta \boldsymbol{x}_l)^T \end{aligned} \tag{B.6}$$

i.e. the matrix C can be considered as a linear combination of $S-1$ matrices D_j which are all constructed from two vectors:

$$\begin{aligned} \boldsymbol{a}_j &= \Delta \boldsymbol{x}_j \\ \boldsymbol{b}_j &= \Delta \boldsymbol{x}_j + \sum_{l\neq i}\Delta \boldsymbol{x}_l \end{aligned} \tag{B.7}$$

according to:

$$D_j = \boldsymbol{a}_j \boldsymbol{b}_j^T \tag{B.8}$$

Clearly, all the columns of D_j are multiples of $\boldsymbol{a}_j$. This means that for the rank R of each D_j:

$$R \leq 1 \tag{B.9}$$

Combining this result with equation B.6 we find that C is a linear combination of $S-1$ matrices D_j each of which has a rank $R \leq 1$. So matrix C cannot have a rank that exceeds $S-1$. The diagonal matrix Λ must have the same rank as C, so Λ cannot have more than $S-1$ nonzero diagonal elements (eigenvalues).

Appendix C

ED analysis of Brownian systems

In chapter 4 it was remarked that the motion of a protein within the essential subspace is believed to be of a diffusive nature [47]. When looking at projections of protein trajectories on the eigenvectors, for instance figs. 3.3, 4.6 and 5.6, a strange phenomenon can be observed: there is a clear periodity along the first few eigenvectors, which becomes less apparent along vectors with smaller eigenvalues.

In order to investigate if this periodicity can result from pure diffusion, we studied a 1000-dimensional Brownian system having coordinates $x_1, x_2 .... x_{1000}$. A trajectory was produced by first setting all coordinate values equal to zero. Then the system was integrated by, during each step, incrementing each coordinate with a random number R, which had a uniform distribution within the interval $[-0.5, 0.5]$:

$$x_m^{n+1} = x_m^n + R \tag{C.1}$$

where x_m^n is the value of coordinate x_m at step n. Random numbers were produced using the RAN function which is implemented in the SGI F77 compiler. Simulations were performed of 1000, 10,000, 100,000 and 1,000,000 steps. From each simulation, a covariance matrix was

constructed and diagonalized. After this, the trajectories were projected onto their first 6 eigenvectors. For the simulations of 1000 and 1,000,000 steps the results are shown in figs. C.1 and C.2. We find very

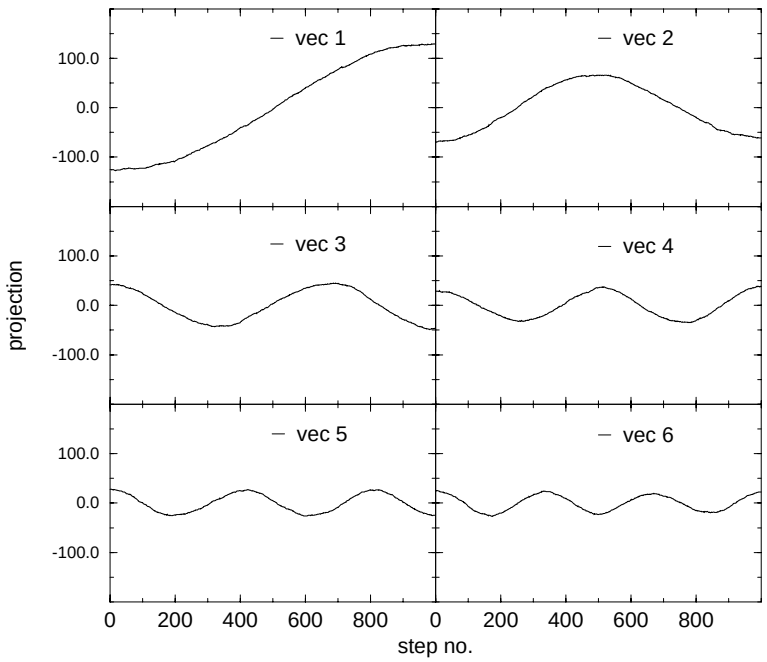

Figure C.1: projections along the first 6 eigenvectors from the simulation of 1000 steps.

clear cosine-like functions. In fact all these projections can accurately be approximated by the relation:

$$p_m^n = \pm A_m cos \frac{\pi n}{mN} \tag{C.2}$$

where p_m^n is the projection at step n onto eigenvector m, N is the total number of steps and A_m is the (positive) amplitude which is related to the eigenvalue λ_m according to:

$$\lambda_m = \frac{1}{N} \sum_{n=1}^{N} (p_m^n)^2 = \frac{1}{N} \sum_{n=1}^{N} (A_m cos \frac{\pi n}{mN})^2 \tag{C.3}$$

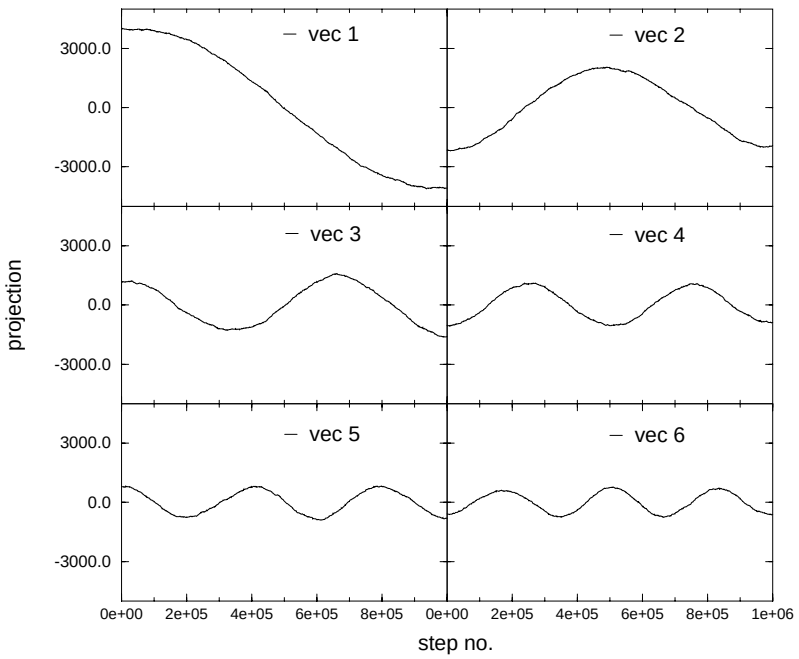

Figure C.2: projections along the first 6 eigenvectors from the simulation of 1,000,000 steps.

giving:

$$A_m = \sqrt{\frac{N\lambda_m}{\sum_{n=1}^{N}(cos\frac{\pi n}{mN})^2}} \tag{C.4}$$

For large N, the summation in the right term may be approximated by an integral, which results in:

$$A_m = \sqrt{2\lambda_m} \tag{C.5}$$

We also compared the eigenvalue curves. Before doing so, they were all scaled according to:

$$\lambda_k^{scaled} = \frac{\lambda_k}{\lambda_1} \tag{C.6}$$

The first 5 eigenvalues obtained in this way, from all simulations mentioned above, are shown in fig C.3. Apparently, the ratio between differ-

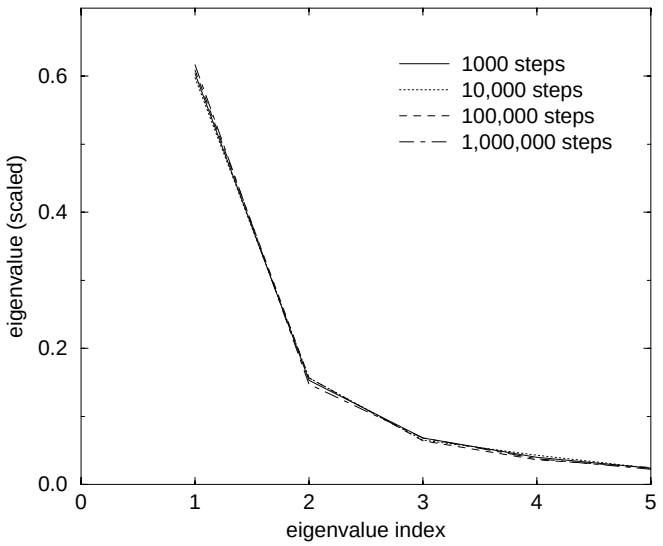

Figure C.3: Scaled eigenvalue curves obtained from Brownian systems of 1000 dimensions (see text).

ent eigenvalues from the same simulation is independent of the length of that simulation. We also checked whether there is a dependency of these ratios on the dimension of the system. We performed four additional simulations of 1000 steps of systems having respectively 200, 400, 600 and 800 dimensions. The scaled eigenvalue curves are shown in fig C.4. Also here, the ratios remain essentially unchanged. Finally we compared the curves with the scaled eigenvalues obtained from the protein simulations of lysozyme and dehalogenase (described in chapters 3,4 and 5) with the 1000 steps simulation of the 1000 dimensional Brownian system. The results are shown in fig C.5. Now clear differences can be seen. This might mean that the essential motion of proteins is of a different nature.

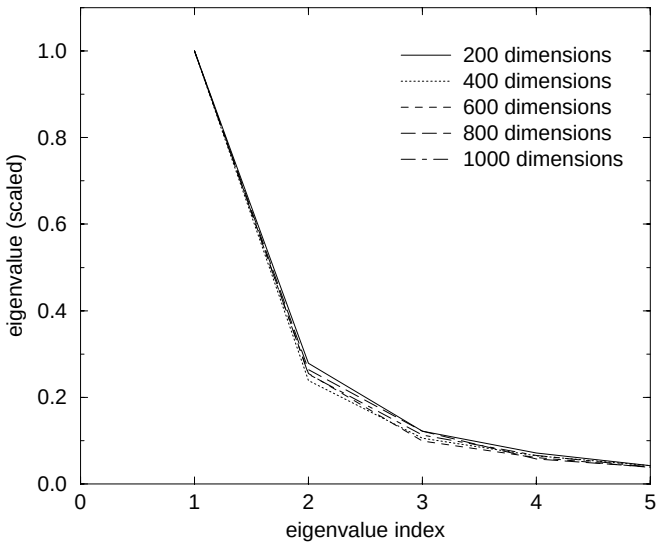

Figure C.4: Scaled eigenvalue curves obtained from 1000 steps simulations of Browian systems with different dimensionality (see text)

However, if diffusive behaviour only takes place in the low-dimensional essential space, this low dimensionality may cause a lack of good sampling. This can be checked by performing a low-dimensional Brownian simulation. The result of such a simulation, of 5 dimensions and 1000 steps is also shown in fig C.5. Also here we see a large deviation from the ideal situation. In fig C.6 the projection of this trajectory onto its first eigenvector is shown. The behaviour is irregular, similar to the projections of the protein trajectories (figs. 3.3, 4.6 and 5.6). It must also be noted that, in contrast to the Brownian systems presented here, the essential space of proteins must be bounded. If during a simulation a boundary along an eigenvector has been encountered, this will affect the mean square displacement along that vector and thereby affect the eigenvalue.

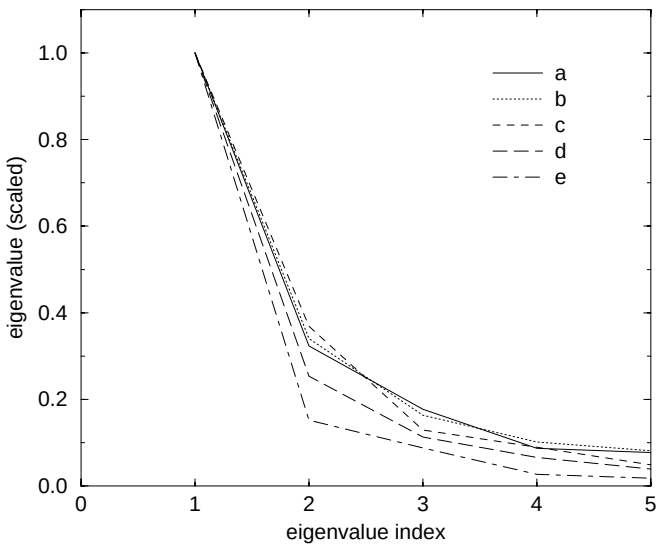

Figure C.5: Scaled eigenvalue curved from simulations of a) lysozyme (chapter 3), b) dehalogenase (chapter 4), c) dehalogenase with Cl^- in the active site (chapter 5), d) Brownian system of 1000 dimensions simulated for 1000 steps. e) Brownian system of 5 dimensions simulated for 1000 steps. Eigenvalues from all proteins were calculated from C_α trajectories.

So far we have not been able to find a mathematical reason for the peculiar periodic behaviour. Still, some remarks can be made. Let's assume that we have simulated an arbitrary system (not necessarily diffusive) during a time interval $[0, T]$ and we have diagonalized the covariance matrix of this complete trajectory. By definition, the cross correlation between projections along different eigenvectors k and l must vanish:

$$\int_0^T p_k(t) p_l(t)\ dt = 0 \quad \text{if} \quad k \neq l \tag{C.7}$$

We also have the requirement that the trajectory, as well as all of its projections, is continuous. Both conditions can only be satisfied if both

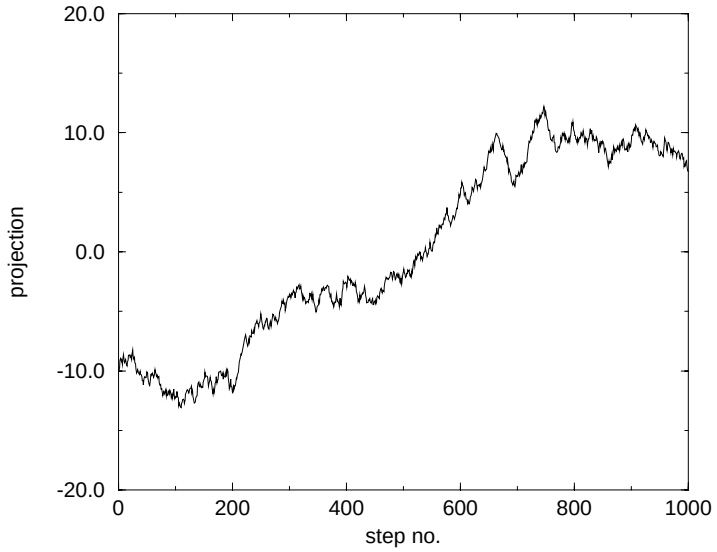

Figure C.6: Projection onto the first eigenvector of a 5 dimensional Brownian trajectory of 1000 steps.

projections have an unequal number of nodes. This does however not account for the observation that the n^{th} projection has n nodes and that the ratio between eigenvalues appears to be independent of the dimension of the system and of the length of the simulation. But it is also possible to give a mathematical expression that predicts the ratio between eigenvalues.

If we consider a continuous diffusive system, eq. C.2 changes into:

$$p_m(t) = A_m \cos \frac{\pi t}{mT} \tag{C.8}$$

If we assume that the equipartition theorem of kinetic energy is valid for this system and we take all masses equal to unity, we must have that along each cartesian coordinate the mean square velocity must be equal, i.e. $\langle v_i^2 \rangle = \langle v^2 \rangle$ for each i. The velocity $\dot{p}(t)$ at time t along an arbritrary

normalized eigenvector $\boldsymbol{\mu}$ given as:

$$\dot{p}(t) = \boldsymbol{\mu}^T \boldsymbol{v}(t) = \sum_i \mu_i v_i(t) \tag{C.9}$$

For the mean square velocity $\langle \dot{p}^2 \rangle$ we find:

$$\begin{aligned} \langle \dot{p}^2 \rangle &= \left\langle \left(\sum_i \mu_i v_i \right) \left(\sum_j \mu_j v_j \right) \right\rangle && \text{(C.10)} \\ &= \left\langle \sum_i \sum_j \mu_i \mu_j v_i v_j \right\rangle && \text{(C.11)} \\ &= \sum_i \sum_j \mu_i \mu_j \langle v_i v_j \rangle && \text{(C.12)} \end{aligned}$$

where $\langle v_i v_j \rangle = 0$ if $i \neq j$, because velocities along different cartesian coordinates are uncorrelated. Furthermore we have $\langle v_i^2 \rangle = \langle v^2 \rangle$. This results in

$$\begin{aligned} \langle \dot{p}^2 \rangle &= \sum_i \mu_i^2 \langle v^2 \rangle && \text{(C.13)} \\ &= \langle v^2 \rangle \sum_i \mu_i^2 && \text{(C.14)} \\ &= \langle v^2 \rangle && \text{(C.15)} \end{aligned}$$

where we have used the fact that $\boldsymbol{\mu}$ is normalized. Equation C.15 shows that the mean square velocity $\langle \dot{p}_m^2 \rangle$ along eigenvector $\boldsymbol{\mu}_m$ is equal for all m. From eq. C.8 we derive:

$$\dot{p}_m(t) = -\frac{\pi A_m}{mT} \sin\left(\frac{\pi t}{mT}\right) \tag{C.16}$$

giving:

$$\begin{aligned} \langle \dot{p}_m^2 \rangle &= \frac{1}{T} \int_0^T \left(\frac{\pi A_m}{mT}\right)^2 \sin^2\left(\frac{\pi t}{mT}\right) dt && \text{(C.17)} \\ &= \frac{1}{2} \left(\frac{\pi A_m}{mT}\right)^2 && \text{(C.18)} \\ &= \frac{\lambda_m}{m^2} \left(\frac{\pi}{T}\right)^2 && \text{(C.19)} \end{aligned}$$

where we have used eq. C.5. Two eigenvalues corresponding to eigenvectors m and l are now related according to:

$$\frac{\lambda_m}{m^2}\left(\frac{\pi}{T}\right)^2 = \frac{\lambda_l}{l^2}\left(\frac{\pi}{T}\right)^2 \tag{C.20}$$

or:

$$\frac{\lambda_m}{\lambda_l} = \frac{m^2}{l^2} \tag{C.21}$$

From eq. C.21 a scaled eigenvalue curve can be derived. Fig. C.7

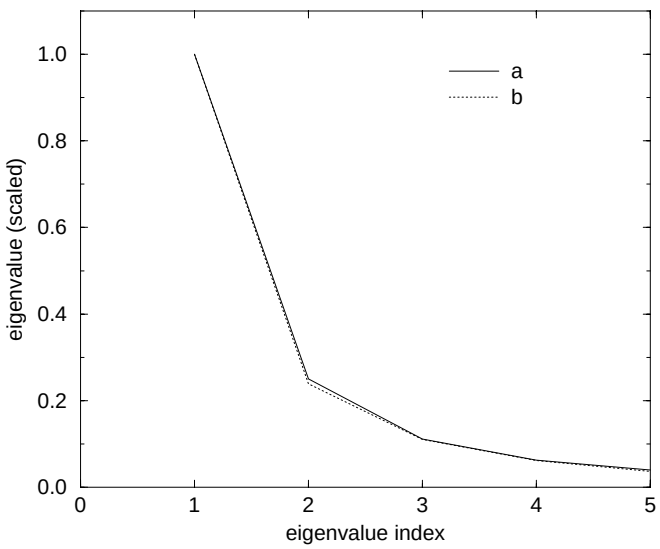

Figure C.7: Scaled eigenvalues. a) derived from the relation $\lambda_m/\lambda_l = m^2/l^2$. b) From the 1000000 steps Brownian simulation of 1000 dimensions.

shows such a curve for a 1000 dimensional system together with the curve from the 1,000,000 steps Brownian simulation mentioned earlier. Both curves are indeed virtually indentical, indicating that the equipartition assumption is valid.

Bibliography

[1] D. B. Janssen, A. Scheper, L. Dijkhuizen, and B. Witholt, Degradation of halogenated aliphatic compounds by X*anthobacter autotrophicus* GJ10, *Appl. Environ. Microbiol.*, **49**, 673-677 (1985).

[2] S. Keuning, D. B. Janssen, and B. Witholt, Purification and characterization of hydrolytic haloalkane dehalogenase from X*anthobacter autotrophicus* GJ10, *J.Bacteriol.*, **163**, 635-639 (1985).

[3] K. H. G. Verschueren, S.M. Franken, H.J. Rozeboom, K. H. Kalk, and B.W. Dijkstra, Refined crystal structures of haloalkane dehalogenase at pH 6.2 and pH 8.2 and implications for the reaction mechanism, *J. Mol. Biol.*, **232**, 856-872 (1993).

[4] A.J. van den Wijngaard, K. van der Kamp, J. van der Ploeg, B. Kazemier, F. Pries, and D. Janssen, Degradation of 1,2-dichloroethane by facultative methylotrophic bacteria, *Appl. Env. Microbiol.*, **58**, 976-983 (1992).

[5] P.J. Kraulis, Molscript: a program to produce both detailed and schematic plots of protein structures, *J. Appl. Crystallogr.*, **24**, 946-950 (1991).

[6] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Essential dynamics of proteins, *Proteins*, **17**, 412-425 (1993).

[7] S. M. Franken, H. J. Rozeboom, K. H. Kalk, and Bauke W. Dijkstra, Crystal structure of haloalkane dehalogenase: an enzyme to detoxify halogenated alkanes, *EMBO J.*, **10**, 1297-1302 (1991).

[8] D.L. Ollis, E. Cheah, M. Cygler, B.W. Dijkstra, F. Frolow, S.M. Franken, M. Harel, S.J. Remington, I. Silman, J. Schrag, J.L. Sussman, K.H.G. Verschueren, and A. Goldman, The α/β-hydrolase fold, *Prot. Eng.*, **5**, 197-211 (1992).

[9] K. H. G. Verschueren, S.M. Franken, H. J. Rozeboom, K. H. Kalk, and B.W. Dijkstra, Crystallographic analysis of the catalytic mechanism of haloalkane dehalogenase, *Nature*, **363**, 693-698 (1993).

[10] F. Pries, J. Kingma, M. Pentenga, G. van Pouderoyen, C.M. Jeronimus-Stratingh, A.P. Bruins, and D.B. Janssen, Site-directed mutagenesis and oxygen isotope incorporation studies of the nucleophilic aspartate of haloalkane dehalogenase, *Biochemistry*, **33**, 1242-1247 (1994).

[11] F. Pries, J. Kingma, G.H. Krooshof, C.M Jeronimus-Stratingh, A.P. Bruins, and D.B. Janssen, Histidine 289 is essential for hydrolysis of the alkyl-enzyme intermediate of haloalkane dehalogenase, *J. Biol. Chem.*, **270**, 10405-10411 (1995).

[12] F. Pries, J. Kingma, and D.B. Janssen, Activation of an Asp124Asn mutant of haloalkane dehalogenase by hydrolytic deamidation of asparagine, *FEBS Lett.*, **358**, 171-174 (95).

[13] C. Kennes, F. Pries, G.H. Krooshof, E. Bokma, J. Kingma, and D.B. Janssen, Replacement of tryptophan residues in haloalkane dehalogenase reduces halide binding and catalytic activity, *Eur. J. Biochem.*, **228**, 403-407 (1995).

[14] J.P. Schanstra and D.B. Janssen, Kinetics of halide release of haloalkane dehalogenase: evidence for a slow conformational change, *Biochemistry*, **35**, 5624-5632 (1996).

[15] W. F. van Gunsteren and H. J. C. Berendsen. *Gromos manual.* BIOMOS, Biomolecular Software, Laboratory of Physical Chemistry, University of Groningen, The Netherlands, 1987.

[16] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints; molecular dynamics of n-alkanes, *J. Comp. Phys.*, **23**, 327-341 (1977).

[17] D.V. Belle, I. Couplet, M. Prevost, and S.J. Wodak, Calculation of electrostatic properties of proteins, *J. Mol. Biol.*, **198**, 721-735 (1987).

[18] P. Ahlström, A. Wallquist, S. Engström, and B. Jönsson, A molecular dynamics study of polarizable water, *Mol. Phys.*, **68**, 563-581 (1989).

[19] P.C. Jordan, J. van Maaren, J. Mavri, D. van der Spoel, and H.J.C. Berendsen, Towards phase transferable potential functions: Methodology and application to nitrogen, *J. Chem. Phys.*, **103**, 2272-2285 (1995).

[20] S.B. Zhu, S. Singh, and G.W. Robinson, Field-perturbed water, *Adv. Chem. Phys.*, **85**, 627-731 (1994).

[21] J. Applequist, J.R. Carl, and K. Fung, An atom dipole interaction model for molecular polarizability. Application to polyatomic molecules and determination of atom polarizabilities, *J. Am. Chem. Soc.*, **94**, 2952-2960 (1972).

[22] *CRC Handbook of Chemistry and Physics*, CRC Press Inc., 67 edition, 1986.

[23] S. Hayward, A. Kitao, and H.J.C. Berendsen, Model-free methods of analyzing domain motions in proteins from simulation: A comparison of normal mode analysis and molecular dynamics simulation of lysozyme, *Proteins*, **27**, 425-437 (1997).

[24] S. Hayward and H.J.C. Berendsen, Systematic analysis of domain motions in proteins from conformational change: New results on citrate synthase and T4 lysozyme, *Proteins*, **30**, 144-154 (1998).

[25] M.O. Hill, Correspondence analysis: A neglected multivariate method, *Appl. Statist.*, **23**, 340-354 (1974).

[26] L. Holm and C. Sander, Parser for protein folding units, *Proteins*, **19**, 256-268 (1994).

[27] T. Ichiye and M. Karplus, Collective motions in proteins; a covariance analysis of atomic fluctuations in molecular dynamics and normal modes simulations, *Proteins*, **11**, 205-217 (1991).

[28] T. Horiuchi and N. Gō, Projection of Monte Carlo and molecular dynamics trajectories onto the normal mode axes: human lysozyme, *Proteins*, **10**, 106-116 (1991).

[29] M.M. Teeter and A.D. Case, Harmonic and quasiharmonic descriptions of crambin, *J. Phys. Chem.*, **94**, 8091-8097 (1990).

[30] D. Perahia, R.M. Levy, and M. Karplus, Motions of an α-helical peptide: Comparison of molecular and harmonic dynamics, *Biopolymers*, **29**, 645-677 (1990).

[31] A. Kitao, F. Hirata, and N. Gō, The effects of solvent on the conformation and collective motions of proteins: normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum, *Chem. Phys.*, **158**, 447-472 (1991).

[32] A. E. Garcia, Large-amplitude nonlinear motions in proteins, *Phys. Rev. Lett.*, **68**, 2696-2699 (1992).

[33] O. Edholm and H.J.C. Berendsen, Entropy estimation from simulations of non-diffusive systems, *Mol. Phys.*, **51**, 1011-1028 (1984).

[34] A. di Nola, H.J.C. Berendsen, and O. Edholm, Free energy determination of polypeptide conformations generated by molecular dynamics, *Macromolecules*, **17**, 2044-2050 (1984).

[35] M. Karplus and J.N. Kushick, Method for estimating the configurational entropy of macromolecules, *Macromolecules*, **14**, 325-332 (1981).

[36] F.C. Bernstein, T.G. Koetzle, G.J.B. Williams, E.F. Meyer Jr., M.D. Brice, J.R. Rogers, O. Kennard, T. Shimanouchi, and M. Tasumi., The Protein Data Bank: a computer-based archival file for macromolecular structures, *J. Mol. Biol.*, **112**, 535-542 (1977).

[37] H. J. C. Berendsen, J. P. M. Postma, A. DiNola, and J. R. Haak, Molecular dynamics with coupling to an external bath, *J. Chem. Phys.*, **81**, 3684-3690 (1984).

[38] P.E. Smith, R.M. Brunne, A.E. Mark, and W.F. van Gunsteren, Dielectric properties of trypsin inhibitor and lysozyme calculated from molecular dynamics simulations, *J. Phys. Chem.*, **97**, 2009-2014 (1993).

[39] A.D. McLachlan, Gene duplications in the structural evolution of chymotrypsin, *J. Mol. Biol.*, **128**, 49-79 (1979).

[40] W.H. Flannery, B.P. Teukolsky, and S.A. Vettering, *Numerical Recipes*, Cambridge University Press, 2nd edition, 1987.

[41] L. Stryer, *Biochemistry*, W. H. Freeman and Co., New York, 3th edition, 1988.

[42] A. R. van Buuren, S. J. Marrink, and H. J. C. Berendsen, *J. Phys. Chem.*, **97**, 9206-9212 (1993).

[43] D. van der Spoel, A.R. van Buuren, D.P. Tieleman, and H.J.C. Berendsen, Molecular dynamics simulations of peptides from BPTI: A closer look at amide-aromatic interactions, *J. Biomol. NMR*, **8**, 229-238 (1996).

[44] H.J. Rozeboom, J. Kingma, D.B. Janssen, and B.W. Dijkstra, Crystallization of haloalkane dehalogenase from X*anthobacter autotrophicus* GJ10, *J. Mol. Biol.*, **200**, 611-612 (1988).

[45] D. van der Spoel, H. J. C. Berendsen, A. R. van Buuren, E. Apol, P. J. Meulenhoff, A. L. T. M. Sijbers, and R. van Drunen. *Gromacs User Manual.* Nijenborgh 4, 9747 AG Groningen, The Netherlands. Internet: `http://rugmd0.chem.rug.nl/~gmx`, 1996.

[46] W. Kabsch and C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, **22**, 2577-2637 (1983).

[47] B.L. de Groot, A. Amadei, R.M. Scheek, N.A.J. van Nuland, and H.J.C. Berendsen, An extended sampling of the configurational space of HPr from E. coli, *Proteins*, **26**, 314-322 (1996).

[48] G. Vriend, What if: a molecular modeling and drug design program, *J. Mol. Graph.*, **8**, 52-56 (1990).

[49] P. Grochulski, Y. Li, J.D. Schrag, and M. Cygler, Two conformational states of C*andida rugosa* lipase, *Prot. Sci.*, **3**, 82-91 (1994).

[50] D.E. Stewart, A. Sarkar, and J.E. Wampler, Occurrence and role of cis peptide bonds in protein structures, *J. Mol. Biol.*, **214**, 253-260 (1990).